Validation of Long-Term and Post-Acute Care Quality Indicators

CMS Contract No:
500-95-0062/
T.O. #4

Final Draft Report

August 2, 2002

Prepared for
Yael Harris, Project Officer
Centers for Medicare and Medicaid Services
Office of Clinical Standards and Quality / S3-02-01
7500 Security Boulevard
Baltimore, Maryland 21244

Principal Authors
John N. Morris
Terry Moore
Rich Jones
Vincent Mor
Joseph Angelelli
Katherine Berg
Christine Hale
Shirley Morris
Katharine M. Murphy
Melissa Rennison

Contract Agencies:
Abt Associates Inc. (*Prime Contractor)*
55 Wheeler Street
Cambridge, Massachusetts 02138-1168

HRCA Research and Training Institute
1200 Center Street
Roslindale, Massachusetts 02131

Brown University
Center for Gerontology and Health Care Research
171 Meeting Street, Box G-B213
Providence, Rhode Island 02912

Key Staff:
John N. Morris
Principal Investigator, HRCA
Katharine M. Murphy
Co-Investigator, HRCA
Vincent Mor
Co-Investigator, Brown University
Katherine Berg
Co-Investigator, Brown University
Terry Moore
Project Manager, Abt Associates

CMS Project Officer:
Yael Harris

Internal Review

Project Director

Technical Reviewer

Management Reviewer

Contents

Appendices

Executive Summary

Assessments of health care quality and the dissemination of resulting information about quality is becoming more widespread in the U.S. These assessments are most frequently in the form of "quality indicators" that are intended to reflect the quality of the care delivered or the patient care outcomes that can be attributed to the care delivered by various health care providers. In this report, we summarize the results of our efforts to validate a series of quality indicators for use with chronic and post-acute care nursing home residents. Some of the other sources of quality indicators that are in current use include the Agency for Health Care Quality's Inpatient and Prevention Quality Indicators, the CAHPS (Consumer Assessment of Health Plans), the National Committee for Quality Assurance's HEDIS measures, Outcome-based Quality Indicators (OBQIs) for home health, dialysis care quality measures, and nursing facility quality indicators. The development of the latter three types of measures have been funded by the Centers for Medicare and Medicaid Services (CMS) in whole or in part, and federal initiatives are underway to utilize the home health and nursing facility measures for regulatory as well as public reporting purposes. Dialysis care measures are currently publicly reported on the CMS website.

The types of measurement information commonly utilized in making judgments about the value of a particular quality indicator include whether the measure has face (or clinical) validity and construct validity, and whether it reliably captures and measures what it purports to measure (validity). Earlier work under this contract (the CMS-sponsored "Development and Validation of Long-term and Post-acute Care Quality Indictors" project) established a set of 45 Minimum Data Set-based (MDS) quality indicators for use in nursing facilities that fulfilled select measurement criteria such as those cited above. These indicators were provisionally recommended for use by CMS, pending an assessment of their reliability and validity (Abt Associates Inc., Oct-2001).

This report summarizes work performed to date to validate these 45 existing and newly developed quality indicators for the long-term and post-acute care populations residing in nursing facilities. Thirty indicators applicable to the chronic (or long-term) care population that were originally developed by others were evaluated, as were 15 newly developed measures for the chronic and post-acute care populations[1]. To our knowledge, the only previous work done to validate any nursing home quality indicators of this type was performed by the Centers for Health Systems Research and Analysis at the University of Wisconsin (CHSRA)[2] (see Zimmerman et al., 1995; Zimmerman et al., 1999; and Zimmerman and Karon, 1997). The list of measures examined in this study may be found in Tables 1 and 2.

Many of the indicators studied here are already in use by CMS in the quality monitoring system utilized in the long-term care survey process. Many facilities actively use these measures for enhancing internal quality performance. In addition, nine of the measures reported upon here are

[1] Original developers of existing quality indicators examined in this report include the Centers for Health Systems Research and Analysis, University of Wisconsin, LTCQ Inc., and J.D. Ramsey.

[2] "Validate" in this context means to clinically review the indicator against medical record and other primary data. Other developers may have performed other types of validation, for example, through secondary data analysis or the convening of industry experts, but this is not the type of validation done in this study nor in the CHSRA validation studies.

being publicly reported for the states of Colorado, Florida, Maryland, Ohio, Rhode Island, and Washington as part of the CMS "nursing home quality initiative" (NHQI)[3].

Defining Nursing Home Quality

Nursing facility quality is multidimensional, encompassing clinical, functional, psychosocial and other aspects of resident health and well being. In this examination of nursing facility quality indicators, all listed aspects of resident functioning are addressed in varying degrees, and the needs of chronic residents and post-acute patients are separately examined. In most cases, multiple quality indicators (QIs) are recommended within a given domain of quality (e.g., clinical quality), and we propose that CMS utilize several QIs from each domain for purposes of public reporting, quality monitoring, and performance improvement. As stated previously, quality of care is necessarily multidimensional. No single QI is likely to capture overall facility quality. Facilities may perform extremely well on one type of QI, but may not perform nearly as well on another. Indeed, two papers recently confirmed this hypothesis, one using New York state data and the other data from Massachusetts (Mukamel and Brower, 1998; Porell and Caro, 1998). It is therefore important to present different indicators across multiple domains for a full view of facility quality performance.

Measurement of Quality

The research design utilized in this quality indicator validation study follows that of other researchers who have concluded that quality must be measured by examining the interaction of structural, process and outcome measures (Donabedian, 1980; Sainfort et al, 1995; Zimmerman et al, 1995; Ramsey, et al, 1995). Each of these quality dimensions was incorporated into our hypotheses concerning the factors that enable a facility to prevent clinical and other problems from occurring, our subsequent collection of data from nursing facilities, and the analyses of these data.

Validation Study Parameters

The final analytic sample for this study was comprised of 209 freestanding and hospital-based facilities located in six states: California, Illinois, Missouri, Ohio, Pennsylvania and Tennessee. Facilities were selected for participation in the study based upon their quality indicator scores (observed over the prior year) on twenty QIs, their geographic location and their willingness to participate in the data collection protocols. The total patient sample included in our on-site field review comprised some 5,758 chronic and post-acute patients, although these facilities serve over 20,000 residents at any one point in time. Compared to all facilities in the states from which they were selected, participating facilities tended to be somewhat larger, were more likely to be non-profit and were less likely to be located in rural settings.

Both resident-level and facility-level data were collected in each sampled facility from November 2001 through June 2002. At the resident level, medical records were reviewed to determine care processes provided to a representative resident sample during the time period of interest in twenty-one quality dimensions, such as physical restraint use, pressure ulcers and pain. The types of care processes reviewed included whether comprehensive assessments other than the MDS were performed, whether physicians were notified in a timely manner following resident change in status, and whether care planning was documented in the record for identified problems. In addition, a subset

[3] These are designated as "pilot" quality indicators and are listed in Tables 1 and 2.

of MDS items was independently assessed by research nurses for later comparison to facility MDS assessments. Facility-level data collected included an administrative survey in which questions were asked of the administrator and director of nursing, and an observation of the general facility environment.

Methods

Description of the Quality Indicators

In constructing the set of quality measures evaluated here, there has been a concern for possible inter-facility variation in the types of residents admitted and served by nursing facilities; difference in the mix of residents served across facilities raises the possibility that inter-facility comparisons may be biased. To control for this possibility, where deemed necessary, three adjustment strategies have been applied.

1) For all of the indicators a denominator exclusion rule was applied (e.g., residents near death). These residents were not considered in the calculation of the quality indicator.
2) For four of the CHSRA indicators, two sub-versions of the same overall indicator were created for each facility, one applying to high-risk residents, the other to low risk residents. In addition, an overall high/low risk indicator was calculated.
3) For many of the other indicators, including those created by the project team and LTCQ Inc., some type of statistical regression-based covariate adjustment strategy was employed. For many indicators, this involved traditional resident-level covariates, supplemented in many instances by a new type of facility-based adjustment based upon resident characteristics upon admission. This is referred to as the "facility admission profile". QIs constructed using a facility admission profile are designated as such in Tables 1 and 2.

Testing the Reliability of the Quality Indicators

In each participating facility research nurses sampled up to 30 residents records, observing and speaking with (if possible) the resident to complete a reduced form version of the MDS in order to allow for a comparison of the MDS based upon facility assessors and that completed by the research nurse assessor. The rationale for examining the reliability of the QI information across all our participating facilities was to allow for the possibility that poor data quality might compromise our ability to adequately test the validity of the QIs. Having information on the average reliability of the MDS data on which the QIs are based allowed us the possibility of excluding facilities with poor data quality from the analyses.

Over 100 MDS data elements were incorporated in the reliability study. A Kappa statistic was used to calculate the level of agreement between the facility and research nurse assessor. This statistic is more stringent than merely calculating the percentage agreement because it adjusts for the possibility of chance agreement that can occur if the condition in question is relatively rare (something true for many of the QIs).

Validation of the Quality Indicators

During the development of data collection tools for this study, expert clinical panels were convened to develop empirically based hypotheses about what constitutes quality of care in a given dimension

(e.g., pain, activities of daily living (ADL)). This effort met with varied success, as there appear to be relatively few well-studied, research-based "standards of care" in use in the nursing facility environment[4]. In cases in which no empirical evidence could drive theories about what components of care qualify a nursing facility as a "good" performer, the expert panels created their own hypotheses. These hypotheses were then utilized to 1) develop data collection instruments to assess nursing home care processes and structures, and 2) direct analysis of these data.

For the primary validation task, individual validation elements, as well as a series of summary scales, from the three data collection tools (Medical Record Review, Administrative Survey, and Environmental Observation) were categorized by quality of care construct (or hypothesis), and then evaluated to determine the degree of their relationship to each quality indicator. The final categorization of quality of care constructs were defined as "preventive" and "responsive".

- ***Preventive*** strategies represent the class of actions that "good" facilities choose to follow in an attempt to minimize the emergence of problems; these strategies are anticipatory in character. Data elements categorized into the preventive construct include staff training, higher staff resource levels, and facility efforts at continuous quality improvement (CQI).

- ***Responsive*** strategies represent actions that facilities are likely to use as they recognize that residents have ongoing or emerging problems in different quality areas. They represent a service response "audit trail," and as such confirm that staff have recognized the problem. Externally, these facilities will be observed to have higher QI scores, but the medical record will reflect a recognition that action must be taken in response to identified resident problems. Examples of data elements gathered on-site that are categorized as responsive are the documentation of comprehensive assessments (other than the MDS), documentation of changes in resident status, and referrals to specialists from inside and outside of the facility (e.g., physicians).

In summary, preventive strategies work to reduce the prevalence or incidence of quality problems measured by the QIs. On the other hand, responsive strategies reflect the fact that quality problems may have emerged in the resident population and as such reflect a "failure" of the facility to prevent the problem (or failure to achieve expected improvement outcomes). Consequently, responsive strategies are associated with an increased prevalence of problems (i.e. quality indicators).

While the constructs created from the various sources of data were conceptualized as falling into one class or another, clinically and administratively relevant data elements thought to be related to particular QIs might have been able to be classified as either preventive or responsive. Thus, our final classification of the validation elements was done based both upon how they related to one another as well as how they related to the QIs. While seeming to represent a "circular" logic (i.e. using one construct to validate another and then to apply the same logic in the other direction), this is a process that characterizes most efforts at construct validation. Thus, the validation elements and constructs were examined for directionality relative to selected QIs and then the QIs were each formally tested against the battery of constructs (classified as preventive or responsive) to determine whether facility records, care processes and structures related to the QIs in the expected direction.

[4] The best examples found of empirically-based nursing facility care practices came from clinical guidelines established by the Agency for Health Research and Quality, such as the Pain Clinical Practice Guidelines.

For each of the constructs or individual data elements categorized as preventive or responsive, the relationship between it and the full array of quality indicators under study was reviewed. To be found acceptable, the construct had to have a consistent relationship across multiple quality indicators. For example, to be classified as preventive, a data element (e.g., a CQI monitoring protocol) had to be related to several quality indicators. We required that there be a clear directional relationship between the construct and the quality indictors. Specifically, preventive elements had to always show a positive relationship to lower (less problematic) QI rates, while responsive elements had to demonstrate a positive relationship to higher (more problematic) QI rates. In other words, the correlation between preventive elements and quality indicators had to be negative, and the correlation between responsive elements and quality indicator scores had to be positive to be considered clearly directional.

In evaluating the validity of the quality indicators, several summary measures were created and reviewed:

1) a count of the number of significant preventive or responsive validation elements for the quality indicator, with the greater the count, the greater the confidence in the relationship;
2) a measure of the pooled association of the list of significant validation elements with the quality indicator. The latter is derived from a regression equation, and in this case represented by the multiple correlation coefficient. This is a multivariate-derived value that resembles a standard bivariate correlation[5]. In reviewing these values, we settled on a combination of two factors in assigning each of the candidate quality indicators to one of three "valid" categories: Top, Mid, and Not Validated; and
3) the underlying reliability of the MDS item and resulting QI.

To understand how these preventive and responsive factors were applied in establishing the validity of a QI, we provide examples of how these elements individually relate to two of the chronic quality indicators, "Pressure ulcer prevalence" (high & low risk) and "ADL worsening". Both indicators are assigned to the Level I, Top Validity category, and both achieved this status on the basis of the preventive elements alone. For the Pressure ulcer indicator, there was also a substantial array of individual responsive relationships, while for the other, ADL worsening, there was only one item of this type.

The ***Pressure ulcer*** indicator quantifies the proportion of at-risk residents in a facility that have a pressure ulcer (i.e., bed sore, decubitus ulcer, pressure sore) of severity ranging from one persistent area of redness that does not disappear when pressure is relieved to one or more open wounds where the full thickness of skin and subcutaneous tissue is lost and underlying bone or muscle is exposed.

There are a large number of clinical and functional risk factors for pressure ulcers (e.g., poor nutrition, incontinence, diabetes, immobility); thus, a number of preventive activities and responsive factors were evaluated. Preventive activities, in general, relate to the handling of at-risk residents and treatment of conditions that contribute to or mitigate pressure ulcer risk. Responsive activities, in general, define actions that a facility's caregivers take to document, communicate and attempt to ameliorate pressure ulcers once present.

[5] Note: this value can be squared to get the classic R2 estimate of explained variance.

Preventive activities for pressure ulcer prevalence included the screening, assessment, and treatment for conditions placing residents at risk for pressure ulcers. Thus, the following individual data elements or constructs were found to be associated with lower pressure ulcer prevalence:

- More frequent scheduling of assessments for suspicious skin areas.
- Weekly routine assessment using a standard protocol for delirium, that would - if present - keep residents bed-bound.
- Observations on the environmental assessment of residents walking or otherwise out of bed.
- Observations on the environmental assessment of caregivers providing assistance to residents with nutritional needs.
- A constructed scale expressing the extent to which a facility manages clinical, psychosocial, and nutritional complications across domains in a manner consistent with high quality care (expressed as a single factor score).

Staffing factors provide additional (albeit indirect) evidence of preventive activities. For example, staffing items related to pressure ulcer prevalence were 1) the absence of facility management change; and 2) the extent that a facility did *not* rely upon floats or contract staff.

Responsive activities for pressure ulcer prevalence include policies, procedures or actions taken by caregivers in response to existing or newly detected pressure ulcers. Identified activities include:

- Comprehensive assessment (other than the MDS) of pressure ulcers documented in the medical record.
- Assessment of pressure ulcers by a physician.
- Clear documentation in the medical record that the resident has a problem in this area or that the resident's condition has changed relative to pressure ulcers.
- Where change was noted in the medical record, there is documentation that this change 1) was evaluated within 72 hours, 2) resulted in a notification to physician or therapist, 3) resulted in a referral to a consultant, and/or 4) resulted in a change in the care plan.

An additional theme related to pressure ulcers was a constructed measure of the extent to which the medical record and care plan agree that pressure ulcers are a problem. This level of agreement signals facilities with a well-integrated system for problem recognition and treatment implementation.

For ***ADL worsening***, there were 17 significant preventive elements and one significant responsive element. From this set of preventive elements, three primary themes emerge: attention to the resident as an individual, an engaging and safe environment, and good continence care. Further explanation of these themes and related data elements follows.

- Maintaining ADL gains is related to a concern with what the resident is thinking and who he or she may be as a person, as seen in areas related to cognition, behavior, and pain. Better outcomes (i.e. facilities have lower rates of ADL worsening) are observed when there are: 1) CQI monitoring protocols in place for behavioral function and communication; 2) weekly routine screening of communication and pain, using standard protocols; and 3) rooms that are personalized with furniture, photos, and other things from the resident's past.

- Maintenance of ADLs is also related to an environment in which the resident is up and out of bed and engaged in activities. Better outcomes are related to a series of things that were

observed by the research nurse about the facility, including: 1) residents being up and about; 2) residents seen to be walking or independently moving about the facility with or without assistive devices; and 3) indications that a variety of activities are available for residents with different capabilities. Related data elements observed during inspection of the facility environment were that public and common areas were well lighted and resident safety had been considered.

- Finally, there was a link to facility efforts aimed at good continence care. Preventive elements relating to this theme include: 1) a scale that counted up to 15 "good" incontinence management items; 2) a scale that focused on care practices relevant to promoting improved levels of continence; 3) a scale that looked specifically at ADL training approaches that were targeted to helping residents maintain continence patterns; and 4) a CQI monitoring protocol in place for bladder incontinence.

Examination of the Performance of the Facility Admission Profile

In addition to evaluating the validity of each of the 45 QIs, two sets of analyses were conducted to examine the performance of the proposed risk adjustment approach earlier recommended by this project team. Each studied different aspects of the facility-level adjustment mechanism (referred to as the facility admission profile, or FAP).

- First, we compared the validity of raw, or non FAP-adjusted, quality indicators to the validity of FAP-adjusted indicators.
- Second, we tested the impact of systematic measurement bias on quality indicators, as described below.

Findings

Reliability was evaluated in several ways. Research nurse MDS assessments were compared to facility-generated MDS assessments to generate the following statistics: 1) percent agreement between "gold" standard nurses and facility nurses; 2) MDS item-level Kappas; and 3) Kappas for a subset of the QI where these could be established (i.e., for prevalence QIs only).

Table 1 displays reliability and distributional statistics for each of the quality indicators for the 209 facilities in the national study sample. Reliability was assessed using the weighted Kappa statistic, with a value of .40 or higher being considered indicative of inter-assessor agreement, while a value of .75 or higher is indicative of superior inter-assessor reliability. In this case the weighted Kappas reflect the cross-sectional reliability of the MDS items that comprise the numerator of the quality indicator (e.g., the numerator for the "Falls" QI is MDS item J4a). Using this standard, only one of the MDS items for a QI numerator falls below the .40 threshold (MDS item N2, which makes up the "Little to no activity" QI). Thirty-one of the quality indicators are based on MDS items with an average weighted Kappa of .70 or higher.

Table 1 also displays the mean rates of the quality indicators across the 209 sampled facilities. As seen here, only two quality indicators have very low prevalence (i.e. < five percent). The rate of the chronic care "New insertion of indwelling catheter" indicator is two percent, and the rate of the post-acute care "Failure to improve and manage delirium" indicator is three percent across the sampled facilities. Five of these QIs have very high prevalence (i.e., > 60 percent). The rate of "Bladder and

bowel incontinence – high and low risk" is 62 percent, the rate of "Bladder and bowel continence – high risk" is 93 percent. Similarly, the chronic care "Improvement in walking" indicator is 82 percent. Two post-acute care indicators, "Failure to improve during the early post-acute period" and "Failure to improve or prevent respiratory problems" have rates of 63 and 92 percent, respectively. The rate of occurrence of various QIs is another criterion that should be taken into consideration when evaluating the utility of various QIs, as extreme skews in the rates of occurrence may indicate QI instability, as well as poor utility in detecting inter-facility variation.

The validity of the FAP-adjusted quality indicators was examined. Non FAP-adjusted and FAP-adjusted quality indicators were equally valid in all but eight instances. In four, two of which ("Improvement in walking – PAC" and "Inadequate pain management – PAC") are currently in the CMS Nursing Home Quality Initiative pilot project, validity was higher for the FAP-adjusted measures. For the other four, validity for the FAP-adjusted measures was lower. The FAP models did not out-perform the non-FAP models; they did not provide scores that were systematically superior.

Finally, an exploration of the presence of "measurement bias" was also completed, in order to understand whether particular QIs are more subject to over- or under-reporting by facilities than others. If this were the case, we would be able to evaluate the ability of the facility admission profile to capture this measurement bias. By and large, while there was inter-state variation in the extent of over or under-reporting, relatively few QIs were observed to have large levels of under or over-reporting in general and relatively few facilities were systematically over or under-reporting the prevalence of quality problems as measured by a multiplicity of QIs. Thus, there is no evidence of systematic bias in facility reporting of the set of prevalence-based QIs evaluated here. The FAP method of risk adjustment therefore cannot be considered an adequate and robust measure of ascertainment bias.

Table 2 displays the summary measures of quality indicator validity. Of the master list of 45 quality indicators, two could not be evaluated due to missing quality indicator data.[6] Thirteen of the chronic care indicators and four of the post-acute care indicators were judged to be in the Level I (Top) validation category. An additional group of sixteen chronic and two post-acute indicators were also accepted as valid, and placed into Level II, the Mid-Valid Category. A total of seven chronic care indicators and one post-acute care indicator were judged not to be valid.

Conclusions and Recommendations

In this national validation study, there is strong evidence that many of the set of 45 reviewed quality indicators capture meaningful aspects of nursing facility performance. We highly recommend for use by CMS and nursing facilities any of the QIs that fall into the Level I validation category, as these QIs have the strongest degree of evidence that they represent real care processes in nursing facilities. The chronic care quality indicators with the highest level of validity include:

- Prevalence of indwelling catheter;
- Bladder/bowel incontinence (high and low risk, high risk, low risk);
- Urinary tract infections;

[6] High and low risk pressure ulcers will be evaluated separately and findings submitted upon delivery of the final validation report.

- Infections;
- Inadequate pain management;
- Pressure ulcers (high and low risk);
- Late-loss ADL worsening;
- ADL worsening;
- Locomotion worsening;
- Improvement in walking; and
- Worsening bladder continence.

Four post-acute care quality indicators are highly valid, including:

- Failure to improve and manage delirium[7];
- Inadequate pain management;
- Failure to improve during early post-acute period; and
- Improvement in walking.

The chronic quality indicators that we recommend rejecting for further use at this time are:

- Behavior symptoms (high risk and low risk);
- Weight loss;
- Antipsychotic use (high risk and low risk);
- Worsening behavior; and
- Worsening pressure ulcers.

The post-acute care indicator that proved not to be valid is "Failure to Prevent or Improve Pressure Ulcers" and therefore should be rejected for use by CMS.

Those QIs that fall into the Level II – Mid Valid category are deemed appropriate for use in measuring nursing facility quality, as they do offer evidence of validity; they are simply not as highly recommended to CMS as those QIs falling into the "Top" (Level I) validation category. In making final determinations about the utility of these QIs for performance improvement, public reporting or other purposes, CMS may want to review both the prevalence and the reliability of these indicators.

A special note is warranted on the "Little or No Activity" quality indicator. While based on the validation effort it was judged to fall into the Mid-Valid (Level II) category, the MDS item on which the indicator is based was found to have poor reliability. Should CMS choose to utilize this indicator for public reporting, facilities will need instruction on proper coding of this assessment item.

In addition to determining which of these sets of nursing facility quality indicators are "valid", or reflecting the care outcomes and issues they are purported to reflect, these results provide evidence that quality indicators measure aspects of care quality that may be amenable to modification through facility practice. For example, facility staffing and policies, practices or procedures are found to be

[7] Again, this QI has a very low rate of occurrence (three percent) in our study sample. The national distribution of this and other indicators should be examined as CMS makes a final determination as to each QI's overall utility.

related to resident quality outcomes and therefore may be modified by facilities to enhance quality of care delivery.

With regard to the facility admission profile, we find no reason to continue to support the universal application of the FAP as currently operationalized. Nonetheless, our analyses also suggest that there are very real inter-facility differences in the mix of residents admitted and who remain to be served by the facility and that these differences are related to the distribution of facilities as measured by the non FAP-adjusted QIs as well as those relying only upon resident-level adjustment. Thus, additional research focusing on the testing of alternate resident and facility level adjustment variables is needed.

Table 1

QI Rates and Weighted Kappas

Quality Indicator	QI Proportional Rate – The Average Across Facilities	Standard Deviation of the QI Rate	The Rate in the Facility with the Lowest Proportional Problem	The Rate in the Facility with the Highest Proportional Problem	Average Weighted Kappa for MDS Items Composing the QI [1]
Chronic Prevalence					
++Behavior symptoms (high&low risk)BEH1	.20	.10	.00	.68	.71
++Behavior symptoms (high risk) BEH2	.23	.11	.00	.69	.71
++Behavior symptoms (low risk) BEH3	.07	.05	.00	.23	.71
Little or no activity SOC2	.12	.12	.00	.77	.28
Prevalence of indwelling catheter CAT2	.07	.05	.00	.32	.71
++Bladder/bowel incontinence (high&low risk) CNT1	.62	.13	.14	.89	.88
++Bladder/bowel incontinence (high risk) CNT5	.93	.05	.76	.99	.88
++Bladder/bowel incontinence (low risk) CNT6	.49	.13	.12	.83	.88
Urinary tract infections CNT4	.08	.05	.00	.31	.53
Falls FAL1	.08	.04	.00	.24	.52
++Infections (pilot) INFX	**.17**	**.08**	**.00**	**.43**	**.50**
++Feeding Tubes NUT1	.08	.05	.00	.27	.80
++Low Body Mass Index BMIX	.12	.05	.00	.31	.85
++Weight loss (pilot) WGT1	**.08**	**.04**	**.00**	**.26**	**.42**
++Inadequate Pain	**.11**	**.08**	**.00**	**.48**	**.73**

Table 1

QI Rates and Weighted Kappas

Quality Indicator	QI Proportional Rate – The Average Across Facilities	Standard Deviation of the QI Rate	The Rate in the Facility with the Lowest Proportional Problem	The Rate in the Facility with the Highest Proportional Problem	Average Weighted Kappa for MDS Items Composing the QI [1]
Management (pilot) PAIX					
++Pressure ulcers (high&low risk) (pilot) PRU1	**.09**	**.05**	**.00**	**.27**	**.74**
++Pressure ulcers (high risk) PRU2	*	*	*	*	*
++Pressure ulcers (low risk) PRU3	*	*	*	*	*
++Burns, skin tears or cuts BURX	.05	.04	.00	.19	.46
Restraints used daily (pilot) RES1	**.07**	**.09**	**.00**	**.49**	**.56**
++Antipsychotic use (high&low risk) (pilot) DRG1	**.21**	**.08**	**.02**	**.43**	**.89**
++Antipsychotic use (high risk) DRG2	.43	.11	.26	.61	.89
++Antipsychotic use (low risk) DRG3	.17	.07	.02	.40	.89
Chronic Incidence					
Late-loss ADL worsening (pilot) ADL1	**.16**	**.09**	**.00**	**.44**	**.84**
ADL worsening ADL2	.08	.07	.00	.33	.83
ADL improvement ADL3	.25	.09	.08	.48	.83
++Locomotion worsening MOB1	.14	.07	.01	.40	.82
++Improvement in walking WALX	.82	.08	.61	.99	.84
++Cognition worsening COG1	.12	.07	.00	.43	.76

Table 1

QI Rates and Weighted Kappas

Quality Indicator	QI Proportional Rate – The Average Across Facilities	Standard Deviation of the QI Rate	The Rate in the Facility with the Lowest Proportional Problem	The Rate in the Facility with the Highest Proportional Problem	Average Weighted Kappa for MDS Items Composing the QI [1]
++Worsening communication COM1	.11	.07	.00	.31	.83
++Delirium DELX	.09	.06	.00	.29	.61
++Worsening behavior BEH4	.07	.05	.00	.24	.72
++Depressed anxious mood worsening MOD3	.15	.07	.00	.37	.60
New insertion of indwelling catheter CAT1	.02	.02	.00	.09	.71
Worsening bowel continence CNT2	.19	.09	.00	.41	.88
++Worsening bladder continence CNT3	.19	.09	.00	.49	.87
++Pain worsening PAN1	.10	.05	.00	.26	.73
++Worsening pressure ulcers PRU4	.07	.04	.00	.27	.74
Post-acute Prevalence					
++Failure to improve and manage delirium (pilot) DELX	**.03**	**.03**	**.00**	**.16**	**.65**
++Inadequate pain management (pilot) PAIX	**.27**	**.10**	**.02**	**.60**	**.72**

Table 1

QI Rates and Weighted Kappas

Quality Indicator	QI Proportional Rate – The Average Across Facilities	Standard Deviation of the QI Rate	The Rate in the Facility with the Lowest Proportional Problem	The Rate in the Facility with the Highest Proportional Problem	Average Weighted Kappa for MDS Items Composing the QI [1]
Post-acute Incidence					
Failure to improve during early post-acute period ADLX	.63	.19	.14	.99	.72
++Failure to improve bladder incontinence CNTX	.55	.09	.32	.79	.73
++Failure to prevent or improve pressure ulcers PRUX	.23	.09	.04	.50	.74
++Failure to improve or prevent respiratory problems RSPX	.92	.05	.77	.99	.53
++Improvement in Walking (pilot) WALX	**.28**	**.14**	**.03**	**.71**	**.77**

Notes:

1 Kappas below 0.4 reflect poor inter-rater reliability; a value between .40 and .60 is indicative of acceptable inter-assessor agreement; and a value of .75 or higher is indicative of superior inter-assessor reliability.

++ Quality indicator was risk-adjusted using facility admission profile.

- Validation analyses were not complete for these QIs

BOLD items indicate measure was using in Nursing Home Quality Initiative Public Reporting Pilot

Table 2

Summary Measures of Quality Indicator Validity

Quality Indicator	Count of Significant Preventive Data Elements[1]	Count of Significant Responsive / Reactive Data Elements	Total Count of Significant Data Elements	Multiple R (Measure of Association) For Preventive Elements	Multiple R For Responsive Elements	Multiple R for All Elements	Degree of Validity [2] I TOP II MID III NOT Valid
Chronic Prevalence							
++Behavior symptoms (high&low risk) BEH1	3	4	7	.34	.31	.43	II
++Behavior symptoms (high risk) BEH2	1	3	4	.25	.30	.39	III
++Behavior symptoms (low risk) BEH3	0	0	0	--	--	--	III
Little or no activity SOC2	8	1	9	.39	.13	.44	II
Prevalence of indwelling catheter CAT2	5	6	11	.45	.71	.78	I
++Bladder/bowel incontinence (high&low risk) CNT1	7	3	10	.50	.45	.66	I
++Bladder/bowel incontinence (high risk) CNT5	8	2	10	.57	.35	.65	I
++Bladder/bowel incontinence (low risk) CNT6	5	3	8	.47	.31	.56	I
Urinary tract infections CNT4	7	8	15	.51	.41	.59	I
Falls FAL1	4	7	11	.27	.40	.50	II
++Infections (pilot) INFX	**6**	**9**	**15**	**.46**	**.36**	**.53**	**I**
++Feeding Tubes NUT1	7	8	15	.44	.40	.54	II
++Low Body Mass Index BMIX	6	1	7	.39	.20	.41	II
++Weight loss (pilot) WGT1	**3**	**0**	**3**	**.27**	**--**	**.27**	**III**
++Inadequate Pain Management (pilot) PAIX	**5**	**4**	**9**	**.32**	**.67**	**.74**	**I**

Table 2

Summary Measures of Quality Indicator Validity

Quality Indicator	Count of Significant Preventive Data Elements[1]	Count of Significant Responsive / Reactive Data Elements	Total Count of Significant Data Elements	Multiple R (Measure of Association) For Preventive Elements	Multiple R For Responsive Elements	Multiple R for All Elements	Degree of Validity [2] I TOP II MID III NOT Valid
++Pressure ulcers (high&low risk) (pilot) PRU1	**10**	**12**	**22**	**.48**	**.43**	**.59**	**I**
++Pressure ulcers (high risk) PRU2	*	*	*	*	*	*	*
++Pressure ulcers (low risk) PRU3	*	*	*	*	*	*	*
++Burns, skin tears or cuts BURX	4	7	11	.30	.34	.47	II
Restraints used daily (pilot) RES1	**3**	**7**	**10**	**.33**	**.48**	**.52**	**II**
++Antipsychotic use (high&low risk) (pilot) DRG1	**5**	**3**	**8**	**.32**	**.31**	**.47**	**II**
++Antipsychotic use (high risk) DRG2	0	1	1	--	.31	.31	III
++Antipsychotic use (low risk) DRG3	1	3	4	.15	.35	.38	III
Chronic Incidence							
Late-loss ADL worsening (pilot) ADL1	**13**	**1**	**14**	**.49**	**.26**	**.51**	**I**
ADL worsening ADL2	17	1	18	.57	.07	.57	I
ADL improvement ADL3	5	0	5	.39	--	.39	II
++Locomotion worsening MOB1	8	1	9	.62	.09	.62	I
++Improvement in walking WALX	9	0	9	.64	--	.64	I
++Cognition worsening COG1	12	8	20	.40	.34	.52	II
++Worsening communication COM1	3	5	8	.29	.31	.41	II
++Delirium DELX	10	0	10	.40	--	.40	II

Table 2

Summary Measures of Quality Indicator Validity

Quality Indicator	Count of Significant Preventive Data Elements[1]	Count of Significant Responsive / Reactive Data Elements	Total Count of Significant Data Elements	Multiple R (Measure of Association) For Preventive Elements	Multiple R For Responsive Elements	Multiple R for All Elements	Degree of Validity [2] I TOP II MID III NOT Valid
++Worsening behavior BEH4	1	1	2	.15	.17	.24	III
++Depressed anxious mood worsening MOD3	7	0	7	.31	--	.31	II
New insertion of indwelling catheter CAT1	8	6	14	.40	.24	.44	II
Worsening bowel continence CNT2	3	1	4	.25	.30	.45	II
++Worsening bladder continence CNT3	6	5	11	.39	.40	.63	I
++Pain worsening PAN1	10	5	15	.37	.40	.51	II
++Worsening pressure ulcers PRU4	3	2	5	.27	.23	.35	III
Post-acute Prevalence [3]							
++Failure to improve and manage delirium (pilot) DELX	**6**	**3**	**9**	**.58**	**.36**	**.62**	**I**
++Inadequate pain management (pilot) PAIX	**5**	**2**	**7**	**.52**	**.36**	**.64**	**I**
Post-acute Incidence							
Failure to improve during early post-acute period ADLX	9	0	9	.59	--	.59	I
++Failure to improve bladder incontinence CNTX	3	0	3	.37	--	.37	II

Table 2

Summary Measures of Quality Indicator Validity

Quality Indicator	Count of Significant Preventive Data Elements[1]	Count of Significant Responsive / Reactive Data Elements	Total Count of Significant Data Elements	Multiple R (Measure of Association) For Preventive Elements	Multiple R For Responsive Elements	Multiple R for All Elements	Degree of Validity [2] I TOP II MID III NOT Valid
++Failure to prevent or improve pressure ulcers PRUX	1	0	1	.12	--	.12	III
++Failure to improve or prevent respiratory problems RSPX	2	0	2	.42	--	.42	II
++Improvement in Walking (pilot) WALX	**4**	**0**	**4**	**.48**	**--**	**.48**	**I**

Notes:

[1] An alpha significance level for the correlation between the validation element and the quality indicator of .09 or lower.

[2] Level I -- Preventive Multiple R Equal to or Greater than .45 – OR -- Total Multiple R equal to or greater than .55

Level II -- Preventive Multiple R Equal to or Greater than .30 – OR -- Total Multiple R equal to or greater than .40

Level III -- Preventive Multiple R Less than .30 – OR -- Total Multiple R less than .40

[3] The sample utilized in evaluation of the post-acute care QIs includes hospital-based transitional care units (TCUs) only [maximum N = 52 facilities]. At the same time, we note that this was one of two analytic samples that could have been used to evaluate the post-acute indicators. Under a second sampling strategy, the TCU sample could be supplemented through the addition of 104 chronic nursing facilities. In each of these facilities there were sufficient numbers of Medicare residents on which to calculate the post-acute quality indicators. Had this second sample approach been the primary strategy to be followed, rather than the TCU approach on which this task rests, the Failure to Prevent or Improve Pressure Ulcer quality indicator would not have been rejected. In fact it would have been placed in Level I, the highest validation category. At the other extreme, had this alternative approach been used, the Improvement in Walking quality indicator would have been placed in Level III, Not Validated.

++ Quality indicator was risk-adjusted using facility admission profile.

* Validation analyses were not complete for these QIs.

--- Indicates that statistics could not be generated due to lack of significant data elements.

BOLD items indicate measure was using in Nursing Home Quality Initiative Public Reporting Pilot

1.0 Background and Overview

1.1 Summary of Project Accomplishments To Date

The "Development and Validation of Long-term and Post-acute Care Quality Indicators" project was intended to assist the Centers for Medicare and Medicaid Services (CMS) in advancing its vision of stimulating quality of care in nursing facilities by developing indicators that reflect clinical and other important care outcomes. This report describes the results of a large-scale validation study designed to reveal whether a select number of quality indicators indeed measure what they are intended to measure. Prior to this validation study, a number of accomplishments took place and are described here, including:

- evaluation of the literature regarding existing quality of care indicators (QIs);
- development of additional QIs (referred to throughout as "MegaQIs") based on areas where there were "gaps" in measurement for long-term (or chronic) and post-acute (or short-term) populations;
- development of a facility-level risk adjustment methodology referred to as the "facility admission profile" (FAP); and
- a pilot study to test data collection strategies.

These project activities are briefly described below, with further information about how the pilot validation study influenced the current study found in Sections 2, 3 and 4.

Review of the literature. The project team conducted an extensive review of published and unpublished literature on all QIs appropriate for use in determining outcomes for long-term and post-acute facility residents/patients. From this review, 143 indicators were identified and evaluated against select criteria. Minimum criteria for selection of QIs for empirical testing was defined as the presence of a clearly specified numerator and denominator, both of which could be operationally defined using Minimum Data Set (MDS) assessment items (See Abt Associates, Oct-2001). Preference was given to QIs that had some form of risk adjustment in order to permit a fairer comparison between facilities with different patient populations (or casemix). Of the 143 indicators identified, 44 indicators were empirically evaluated. After this evaluation, 26 were deemed to have met the Project Team's selection criteria. Alternative forms of the QIs – specifically forms that utilized a facility-level risk adjustor - were then modeled and reevaluated. This process resulted in a final recommendation of 22 QIs for use by CMS.

Development of Additional Quality Indicators. Once the review of existing QIs had been completed, the project team identified gaps in existing QIs where aspects of care were not being sufficiently addressed. Fifty-four additional quality measures were subsequently developed and tested using secondary data. CMS conducted an internal review process of the 54 newly developed QIs (21 chronic, 21 post-acute, and 12 drug therapy indicators) and then stakeholders were given the opportunity to comment on a website where the underlying conceptual framework of the QIs were posted. Comments from stakeholders and industry representatives were useful in informing the modification of the QI definitions. Decisions on the drug QIs were postponed since CMS decided not to go forward with requiring Section U data on the MDS. Organizational QIs were removed from

further consideration. From this review process, 15 newly developed "MegaQIs" were considered for further validation.

Development of a facility admission profile. A main concern in the implementation of an indicator-based quality reporting system is that judgments based on those QIs might be influenced by facility characteristics other than quality of care. The project team investigated the impact of casemix differences resulting from differential admission or discharge practices and of differential ascertainment as the most likely sources for such biased assessments. The results showed that this concern is warranted and that the specification of appropriate risk adjustment models is a key requirement for the validity of any QI. Other analyses conducted revealed that, particularly in smaller facilities, rankings based on QIs may vary substantially over time and, therefore, that statements about QI performance cannot be made with much statistical confidence.

In attempts to capture these differential effects on quality indicator rankings, a series of analyses were conducted, resulting in the development of a new risk adjustment method that incorporates facility admitting characteristics into the construction of QIs. This adjustment method is referred to as the "facility admission profile" (FAP). As further work on this risk adjustment model has been undertaken, the project team has recommended the use of this facility-level adjuster on some but not all QIs. In general, use of the FAP is recommended for QIs where 1) the adjustment model performs well statistically, and 2) the quality dimension in question is one in which it is expected that facilities cannot affect change upon resident admission. For example, facilities with a "restraint-free" philosophy have the ability to limit physical restraint use at resident admission. Thus, no FAP adjustment is recommended for the "prevalence of restraints" quality indicator.

Validation Pilot Study. Data collection instruments were then developed and field-tested in 45 Massachusetts freestanding facilities. The QIs were grouped by eight quality dimensions for validation: the use of inappropriate drugs, falls, pain, delirium, depression, BMI, failure to improve, and pressure ulcers. The pilot study provided a first indication as to which of the hypothesized independent causal measures were related to the facility measures of nursing home quality (Abt Associates, Sep-2001). Pilot study findings are described in Section 2.

1.2 Selection of Measures for Full-scale Validation

While in the field conducting the pilot study (Spring/Summer 2001), CMS embarked upon a public reporting initiative. This initiative called for public reporting of quality indicator data for all nursing facilities in six pilot states (Colorado, Florida, Maryland, Ohio, Rhode Island, and Washington). After this pilot test, CMS plans to expand public reporting of nursing facility quality to all states in the nation. Due to the CMS pilot project initiative and impending national quality reporting, the project team redesigned and expanded the validation strategy to examine all dimensions of quality covered by the final set of quality indicators (n = 45). Rather than directly validate each of the 45 quality indicators, each quality *dimension* reflecting particular quality indicators was examined. For example, data were collected for "undernutrition" as a quality dimension, with the idea that care processes and facility policies collected in this dimension would address the validity of individual measures such as low BMI and weight loss. Table 1.1 presents the set of quality indicators examined and reported on in this Validation Report; in all, 21 dimensions of quality were examined for chronic and post-acute nursing facility patients.

Table 1.1
Set of Quality Indicators Validated

Indicator	Developer
Chronic Care Quality Indicators	
Late-Loss ADL Worsening	CHSRA
ADL Worsening Following Improvement	MEGAQI
ADL Improvement: Improvement in ADLs Among Residents who Exhibited a Capacity for Improvement at the Prior Assessment	MEGAQI
Locomotion Worsening	LTCQ
Maintenance or Improvement in Walking Performance in Persons with Walking Capacity	MEGAQI
Cognition Worsening	LTCQ
Worsening Communication	LTCQ
Delirium: Failure to Prevent New Delirium or Recurrence of Delirium	MEGAQI
Behavior Symptoms Affecting Others	CHSRA
Worsening Behavioral Symptoms	LTCQ
Depressed/Anxious Mood Worsening	LTCQ
Little of no Activity	CHSRA
New Insertion of Indwelling Catheter	LTCQ
Prevalence of Indwelling Catheters	CHSRA
Bladder or Bowel Incontinence Prevalence (high/low risk, high risk, low risk)	CHSRA
Worsening Bowel Continence	LTCQ
Worsening Bladder Continence	LTCQ
Prevalence of Urinary Tract Infections	CHSRA
Falls Prevalence Among Those Without Recent History of Falls	LTCQ
Infections Prevalence	MEGAQI
Prevalence of Feeding Tubes	Ramsey
Low Body Mass Index (BMI) Prevalence	MEGAQI
Weight Loss Prevalence	LTCQ
Pain, Inadequate Management	MEGAQI
Pain, Worsening	LTCQ
Pressure Ulcer (stage 1-4) Prevalence (high/low risk, high risk, low risk)	CHSRA
Worsening Pressure Ulcers	LTCQ
Burns, Skin Tears or Cuts Prevalence	MEGAQI
Restraints (physical) Used Daily, Prevalence	CHSRA
Prevalence of Antipsychotic use in the Absence of Psychotic and Related Conditions (high/low risk, high risk, low risk)	CHSRA
Post Acute Quality Indicators	
Failure to Improve and Manage Delirium Symptoms	MEGAQI
Failure to Improve During Early Post-Acute Period	MEGAQI
Failure to Improve Bladder Incontinence	MEGAQI
Inadequate Pain Management	MEGAQI
Failure to Prevent Pressure Ulcers or Improve Existing Pressure Ulcers	MEGAQI
Failure to Improve or Prevent Respiratory Problems	MEGAQI
Improvement in Walking	MEGAQI

1.3 Overview of this Report

This Validation Report is structured as follows: Section 2 outlines the preliminary pilot study results that influenced the final full-scale validation design. Section 3 explains the data collection process, such as sampling strategy, description of recruitment, and development of data collection tools. In Sections 4 and 5, we describe the methods for the validation of quality indicators. Section 6 contains results from our primary validation findings and reliability and ascertainment bias findings. Section 7 describes preliminary analyses conducted to examine the performance of the facility admission profile. Section 8 contains a discussion of these results, along with conclusions, recommendations, and a description of next steps.

2.0 Summary of Preliminary Pilot Study Results

In February 2001, a pilot study was conducted to test the team's data collection instruments and to provide a provisional analysis of the hypothesized relationship between quality indicator measures and pertinent service input and process measures.

Two samples of data and related data sources were used to test the QIs. Each data sample included MDS-based QIs derived from computerized MDS data and an array of validation elements collected by research staff from participating facilities. The first sample was from an existing data set of 45 facilities owned or managed by the National Health Corporation (NHC). The second data source was obtained under the current CMS "Development and Validation of Long-term and Post-acute Care Indicators" contract from a sample of 45 nursing facilities in Massachusetts (MA).

Data collection protocols were similar for both the existing (NHC) and new primary data collection samples (MA). Staff at each facility, including the Director of Nursing and a representative from Administration, completed self-administered surveys on facility characteristics, care practices, policies, and procedures. In both samples trained research nurses reviewed up to one hundred resident charts per facility. Reviewed records were selected based on computerized algorithms using MDS data, with protocols keyed to specific QI areas -- three in NHC and eight in MA. In addition, facility staff were asked to complete a survey on factual and attitudinal items, and research staff completed a systematic walk-through to characterize the ambience of the nursing home and to observe facility care plan meetings.

During the development of data collection protocols, expert panels had defined hypotheses that linked field data elements to specific QIs. Using these hypotheses as a guide, exploratory data analysis techniques were then used to combine data from staff surveys, medical record reviews, facility "walk-through" surveys, care plan observations and other forms. Pilot results suggested that 29 of the 31 QIs examined pass a minimal threshold of provisional validity. Some QIs appeared to have stronger validity evidence than others. For the seven post-acute care (PAC) QIs, the analyses were suggestive of the indicators being valid. Single data elements from the chronic care sample validated some of the PAC QIs, but many PAC QIs demonstrated validity with multiple scales.

In this preliminary and exploratory study, the team found that aspects of nursing home quality of care could be measured with field survey research instruments. Validation "constructs" or scales derived from these instruments appeared to explain a significant proportion of the variability in nursing home MDS-based quality indicator rates. These results provide preliminary evidence that support the position that MDS-based QIs are valid measures of aspects of care quality provided by nursing facilities. The next step of the team was to test these relationships in a larger, more nationally representative sample of nursing homes. This full-scale validation study is described in the remaining sections of this report.

3.0 Data Collection Process

3.1 Sampling Strategy

3.1.1 Facility Sampling

The original facility sampling strategy was to emphasize facilities at both extremes ("good" performers and "poor" performers) of the observed quality of care continuum in a state. Probable poor facilities were defined as those with a preponderance of "bad" QIs; i.e., they were one standard deviation or more above the state mean for the selected QIs. At the other extreme, probable good facilities were defined as those with a preponderance of "good" QIs; i.e., they were one standard deviation or more below the state mean. Twenty QIs (those with the most promising results from the pilot study) were used to categorize facilities into good and poor performers. The QIs included:

- ADL decline (CHSRA)
- Mobility decline (LTCQ)
- Walking performance (MegaQI)
- Falls increase (LTCQ)
- Cognition worsening (LTCQ)
- Communication worsening (LTCQ)
- Delirium not remitting (MegaQI)
- Behavior high and low risk (CHSRA)
- Behavior worsening (LTCQ)
- Depression new or worse (LTCQ)
- Indwelling urinary catheter (LTCQ)
- Catheter (CHSRA)
- Incontinence high and low risk (CHSRA)
- Infection flare-up (MegaQI)
- Tube feeding (RAMSEY)
- Low BMI (MegaQI)
- Pain poorly managed (MegaQI)
- Pressure ulcer high and low risk (CHSRA)
- Pressure ulcer onset or worsening (LTCQ)
- Anti-psychotic high and low risk (CHSRA)

We then oversampled from the good and bad facilities. As facility recruitment progressed and targeted facilities refused to participate, the "good" vs. "bad" facility dichotomy gave way to some convenience sampling, as it was vital to reach the target sample of 210 facilities. Therefore, facilities in the extreme tails of quality performance are not as concentrated as originally hoped.

In order to optimize recruitment of hospital-based facilities and to obtain a nationally representative sample, six states with large numbers of hospital-based facilities or transitional care units (TCUs) were selected: California, Illinois, Missouri, Ohio, Pennsylvania, and Tennessee. Within each state, the sample of chronic care facilities and TCUs was drawn from contiguous counties with the greatest concentration of TCUs. Long-term care facilities with fewer than 50 beds or with residents with a mean age of less than 50 years were excluded from the sample.

Hospital-based facilities in each state were randomized. As alluded to earlier, some geographic "convenience" sampling was also done due to resource constraints. Table 3.1 illustrates the percentage of facilities from the recruited sample of 219 facilities within each of the four sampling strata.

Table 3.1
Distribution of Facility Sampling Strata by State for 219 Recruited Facilities

	Neutral		Bad		Good		Bad & Good		Total	
State	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
CA	8	8.4	6	15.0	21	27.6	2	25.0	37	16.9
IL	12	12.6	6	15.0	20	26.3	3	37.5	41	18.7
MO	16	16.8	3	7.5	8	10.5	1	12.5	28	12.8
OH	22	23.2	10	7.5	10	13.2	0	0	35	16.0
PA	13	13.7	20	50.0	11	14.5	1	12.5	45	20.5
TN	24	25.3	2	5.0	6	7.9	1	12.5	33	15.1
Total	**95**	**43.4**	**40**	**18.3**	**76**	**34.7**	**8**	**3.7**	**219**	**100**

The six-state sample is distributed as follows: 37 from California, 41 from Illinois, 28 from Missouri, 35 from Ohio, 45 from Pennsylvania, and 33 from Tennessee. For this sample, not every facility had complete data (e.g., for two facilities, the Administrative Survey was not turned in). The resulting available sample size of 209 will therefore serve as the upper limit for the analytical sample. For most comparisons of chronic quality indicators, the available sample was 151 facilities. For PAC QIs, the available sample numbered 166 facilities, 52 of which were TCUs. Compared to all facilities in the states from which they were selected, participating facilities tended to be somewhat larger, were more likely to be non-profit and were less likely to be located in rural settings. Descriptive statistics comparing the study sample to all nursing facilities in the U.S. may be found in Appendix A.

3.1.2 Patient Sampling

A target of 30 chronic residents or post-acute patients was established per facility. In chronic care nursing facilities, the sample was comprised of 10 residents with a recently completed admission MDS assessment; 10 residents with a recently completed quarterly MDS assessment; and 10 residents with a recently completed annual MDS assessment. "Recently completed" assessments were defined as those that were completed no later than one-month prior to the nurse researcher arriving at the site. If a sample could not be captured with recently completed assessments, the nurse assessors looked back as far as 90 days to fulfill the sample. In hospital-based facilities, the sample was the 30 most recently assessed patients.

Table 3.2 presents the distribution of chronic residents and hospital-based patients from the 209 facilities included in the analytical sample. These distributions are further categorized into neutral, bad, good, and bad and good performers. As displayed in the table, a total of 5,758 long-term and post-acute subjects were included in the analytic sample.

Table 3.2
Distribution of Patients by Facility Sampling Strata by State

	Neutral		Bad		Good		Bad & Good		Total	
State	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
CA	190	7.5	155	14.9	428	21.8	41	19.2	814	14.1
IL	350	13.8	173	16.6	566	28.9	86	40.2	1175	20.4
MO	434	17.1	59	5.7	222	11.3	29	13.6	744	12.9
OH	545	21.4	50	4.8	268	13.7	0	0	863	15.0
PA	361	14.2	548	52.6	306	15.6	29	13.6	1244	21.6
TN	662	26.0	57	5.5	170	8.7	29	13.6	918	15.9
Total	**2542**	**44.1**	**1042**	**18.1**	**1960**	**34.0**	**214**	**3.7**	**5758**	**100**

3.2 Description of Recruitment

A recruitment package was mailed to each potential study site. This package included a letter introducing the project team and outlining study procedures, a project overview and fact sheet, and letters of project endorsement from CMS, the American Association of Homes and Services for the Aging, and the American Health Care Association. The package is included here as Appendix B.

HRCA nurse recruiters called each facility within two weeks of the mailing to verify contact information, answer questions, and - when possible - arrange for a site visit by the assessor team. When the first call did not elicit a positive response, nurses continued calling until the contact person agreed to schedule a visit or firmly refused to participate. Recruitment calls averaged six per site. Acceptance and refusal rates are presented in Table 3.3. Refusal rates were higher for chronic care facilities (54.4 percent) than for hospital-based facilities (47.6 percent). Facilities that refused tended to be larger, and a higher proportion were for profit, chain-owned facilities. Reasons cited for refusal to participate are described in Table 3.4.

Table 3.3
Acceptance and Refusal Rates for Chronic Care and Hospital-based facilities

	Chronic care				Hospital-based			
State	**Accepted**		**Refused**		**Accepted**		**Refused**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
California	19	5.6	24	7.1	18	14.5	16	12.9
Illinois	26	7.7	41	12.1	15	12.1	14	11.3
Missouri	26	7.7	30	8.9	2	1.6	5	4.0
Ohio	23	6.8	21	6.2	12	3.5	6	4.9
Pennsylvania	34	10.1	28	8.3	11	8.9	14	11.3
Tennessee	26	7.7	40	11.8	7	5.6	4	3.2
Total	**154**	**45.6**	**184**	**54.4**	**65**	**52.4**	**59**	**47.6**

Table 3.4
Reasons for Refusal by Facility Type

Reason	Chronic care (%) (N=184)	Hospital-based (%) (N=59)	Total (%)
Too busy	40.2	42.1	**40.2**
Staffing	12.5	11.9	**12.3**
Not interested/no reason	30.0	30.5	**30.3**
Corporation refused	10.8	3.4	**9.0**
Other	6.5	13.6	**8.2**

3.3 Development of Data Collection Tools

As the first step in the development of data collection tools for the Pilot Study, the project team convened panels of experts for the eight targeted dimensions of care (pain, pressure ulcer, high risk drugs, body mass index, falls, depression, failure to improve, and delirium) in the spring of 2000. Criteria used in selecting these areas of care included:

- Prevalence of the problem;
- Face validity;
- Availability of treatments for the problem;
- Tendency for facilities to document the problem (so that evidence may be found in the medical record);
- Susceptibility to being able to be verified via other data collection methods (observation, interview); and
- Sufficient variation in the care area.

Expert panel members were comprised of geriatric nurses, physicians and researchers, both from the participating organizations (Abt Associates, Hebrew Rehabilitation Center for the Aged (HRCA), Brown University and the University of Michigan) and from the long-term care industry.

The charge of the panel members was to review the quality dimensions that were to be validated directly and to propose criteria for how to validate them. These dimensions reflected a mix of facility structures (e.g., staffing) and processes (e.g., drug treatment) as well as an array of resident clinical, functional, and psychosocial outcome areas: psychotropic drugs; falls; delirium; depression; undernutrition (body mass index); failure to improve in activities of daily living (ADLs); pain; and pressure ulcers. Each expert panel was asked to develop a data collection protocol intended to distinguish "good" performing nursing facilities from poorer performers in the panel's designated quality dimension (e.g., pain). These data collection protocols would then be used to "validate" the quality measures. Panels were given a template to use as an example, and asked to 1) review all proposed quality measures (chronic or post-acute) relevant to the quality dimension, 2) conduct a brief literature review to ensure panel members were abreast of the latest clinical practice guidelines, standards of care and research in their quality dimension and include the references in their draft protocol, 3) establish a series of hypotheses about what distinguishes a "good" facility from a poor facility in that dimension (e.g., the Pain subcommittee believed that facilities that have a policy in place to guide pain assessment, treatment and evaluation will be more effective in managing and relieving pain), 4) prioritize hypotheses by giving high priority to those that are empirically-based as well as those which the panel believes contribute most to facility quality, and 5) operationalize

hypotheses by developing a series of questions or data items to be gathered to measure facility practices, processes, structures and/or outcomes using sources including medical records, resident observation and interview, staff interview, administrative surveys, environmental observation and family interview. The expert panels were also asked to provide recommendations regarding measurement of global facility practices that might impact upon a facility's provision of quality care in their designated quality dimension (e.g., the Pain subcommittee believed that an interdisciplinary, inclusive care planning process would contribute to adequate pain management).

The effort to have clinical experts specify hypotheses about the essential care processes or structural elements that must be in place in order to label a facility "good" in a particular care domain met with varied success, as there appear to be relatively few well-studied, research-based "standards of care" in use in the nursing facility environment[8]. In cases in which no empirical evidence could drive theories about what components of care qualify a nursing facility as a "good" performer, the expert panels created their own hypotheses.

Review of the material from the dimension-specific panels revealed extensive hypotheses about care processes and structures that comprise "good" facilities in multiple quality domains, but very little in the way of operationalized data collection measures or items. Therefore, the project team further operationalized the recommended concepts and measures into 1) a series of data collection tools that could be completed by trained nurse assessors during medical record review (MRR) and environmental observation and 2) survey instruments designed for completion by facility staff, resident assessment coordinators, Directors of Nursing (DON) and administrators. In many cases these instruments attempt to tap into the congruence of information from different sources. Taking pressure ulcers for example, measures were developed for the DON survey to determine if the facility had written policies and procedures related to risk assessment, prevention and management and follow-up evaluation of pressure ulcers, and if the licensed and non-licensed staff were offered educational programs about pressure ulcers and about facility policies in this area. The MRR sought to determine if there was documentation of risk assessment, preventive measures, and types of interventions and follow-up for residents with pressure ulcers.

An initial version of the data collection instruments was tested in the winter of 2001 for feasibility in two nursing facilities in Rhode Island and Massachusetts. Information gathered during this feasibility study helped the project team to discover potential problems with the data collection tools, such as ambiguous questions, questions that alienate staff or residents and data items that are not recorded as expected. The gained knowledge lead to substantial modifications of the survey tools.

After incorporating changes from the feasibility study, the revised data collection protocols were tested in a pilot study of 45 Massachusetts nursing homes (see Section 2 for a detailed discussion of the pilot study). Nurse researchers found the data collection instruments to be lengthy, and had two major difficulties in completing them fully at each pilot facility: 1) medical records had often been "thinned" for the period of interest, making it time consuming to gather the required medical record data to complete the medical record review tool; and 2) the "QI-specific" resident sampling framework was cumbersome and at times required the nurse researcher to revisit a particular patient's medical record several times in order to complete the MRR. In addition, nurse researchers found it

[8] The best examples found of empirically-based nursing facility care practices came from clinical guidelines established by the Agency for Health Research and Quality, such as the Pain Clinical Practice Guidelines.

difficult to complete the required staff and resident observations *and* the observations of care planning meetings *and* individual staff interviews, while also completing the required medical record reviews.

These nurse researcher experiences, as well as findings from the analysis of the pilot study validation data, led the project team to extensively revise the resident sampling framework, the MRR, the environmental walk-through, and the administrative survey in preparation for the full-scale validation study. The care plan observations were dropped from the data collection protocol, as were the staff interviews and MDS Coordinator questionnaire. Care was taken in these protocol revisions to maintain data elements that reflected the quality dimensions of interest in the pilot (and additional dimensions, cited below). Appendix C contains the final data collection instruments utilized in this study.

A final feasibility test of the new data collection protocols was conducted in a Massachusetts hospital-based facility in late October 2001. This exercise allowed the team to identify time management issues for the nurse assessors and eradicate duplicate lines of questioning. The facility's Director of Nursing was particularly helpful with her questions and concerns regarding the Administrative Survey. These comments, along with other observed difficulties, were used to rework the final set of data collection tools used in the full-scale validation effort.

Full-scale implementation of data collection began in mid-November 2001 and was completed in mid-June 2002. The following is the list of instruments used in this full-scale validation effort:

- Medical Record Review;
- MDS Supplement;
- Administrative Questionnaire; and
- Environmental Walk Through/Resident Observation.

Resident-level Data Collection Tools

Medical Record Review Tool. The purpose of the medical record review (MRR) was to obtain information regarding the care processes and types of patient/resident assessments performed by sampled facilities on select areas. The intent of this tool was to assist the research team in understanding the relationship between a facility's quality indicator rates and its resident-specific care processes. The following 21 care areas (or quality dimensions) were reviewed during the MRR:

- Cognitive Impairment;
- Communication;
- Delirium;
- Depression/Mood;
- Behavior Problems;
- ADL Improvement;
- ADL Decline;
- Mobility/Walking;
- Falls;
- Anti-psychotic Drugs;
- Pain;

- Physical Restraints;
- Feeding Tubes;
- Undernutrition / Low BMI / Weight Loss;
- Indwelling Urinary Catheter;
- Bladder Incontinence;
- Bowel Incontinence;
- Infections;
- Pressure Ulcers / Potential for Skin Breakdown;
- Burns, Abrasions, Skin Tears; and
- Little or No Involvement in Activities.

For each of these domains, nurse assessors reviewed the medical record (including nursing progress notes, physician orders and progress notes, care plans, therapy consults and notes, medication administration records, flow sheets and other interdisciplinary notes and consults) for resident care and status documentation. Specifically, assessors looked for documentation on comprehensive assessments, problems/issues, change in status (within certain time frames), referrals, treatments and nursing care plans. All MRR information was entered into the "MedQuest" computer software program, backed up onto diskette and archived by the nurse assessors.

Supplement MDS Assessments. The "MDS Version 2.0 for Nursing Home Resident Assessment and Care Screening Supplement" was used to conduct assessments on all patients in the sample (see Section 3.1.2 for a description of the resident sample). This assessment contained questions regarding:

- cognitive patterns;
- communication/hearing patterns;
- mood and behavior patterns;
- physical functioning and structural problems;
- continence in last 14 days;
- disease diagnoses;
- health conditions;
- oral/nutritional status;
- skin conditions;
- activity pursuit patterns;
- medications;
- special treatment procedures; and
- discharge potential and overall status.

The nurse assessors used the resident's record, communication with and observation of the resident (when the resident was deemed capable by facility staff of providing an informed consent), communication with direct-care staff (e.g., nursing assistants, activity aides) and communication with licensed professionals (when available) to complete the evaluation. To ensure impartiality, nurse assessors were instructed to complete the supplemental MDS assessment before reviewing the facility's MDS assessment.

Facility-level Data Collection Tools

Administrative Questionnaire. The Administrator questionnaire included questions regarding:

- staff responsibilities;
- staff/resident/family involvement in care;
- resident status;
- access to specialists/consultants;
- clinical communication channels;
- staff turnover;
- staffing ratios;
- planning processes;
- information on the organization; and
- training and orientation of staff.

These areas were selected for two reasons: 1) the expert panels had developed hypotheses regarding the impact of issues such as staff training and communication on quality; and 2) the Project Team agreed that certain facility-level processes and systems (e.g., communication, care planning) are vitally linked to quality outcomes. While there are many communication patterns represented in a nursing facility, the ones that seem to be most critical are those that involve communication of resident status among facility staff and cognizant physicians. Care planning processes are also considered vital to the successful care of residents. These questions were designed to understand the facility processes that facilitate or impede care delivery, and the relationship of these processes to nursing facility quality of care.

Environmental Walk Through/ Resident Observation. The aim of the Environmental Walk Through/ Resident Observation was to gain an overall understanding regarding whether the facility is "resident-centered", what the "feel" of the facility is, and what the nature of staff interactions with residents are. A series of general environmental measures were employed to describe the responsiveness of the milieu to resident strengths, needs, and problems that include general care environment measures (e.g., nature of physical environment, communication strategies, environmental manipulation and resident interactions with staff). These measures were collected through assessment, surveillance, and observation of staff technique. The data collectors on site recorded their observations three times per day at approximately 10:00 a.m., lunchtime and 2:00 p.m. to obtain a comprehensive picture of the facility care environment.

3.4 Description of Nurse Researcher Training Program

3.4.1 Qualifications of Nurse Researchers

The Peer Review Organizations (PROs) in participating states were responsible for hiring field data collectors. Preference in hiring was given to registered nurses with chart review experience who also had experience in a long-term care setting and/or in completing the MDS Version 2.0. Among the final group of hires all but one was an RN, several had both long-term care and MDS experience, and all had PRO experience abstracting data from medical records.

3.4.2 Summary of Training and Certification Program

Prior to initiating the field study in November 2001, we conducted a five-day training and certification program in Cambridge, Massachusetts for the newly hired nurse assessors. The majority of assessors were trained in our data collection procedures during this session. Two additional sessions were held in December and January at HRCA for assessors who were unable to attend the first program. Trainers included the CMS Project Officer, three members of the Project Steering Committee (including two RNs), five experienced RN researchers from HRCA who had participated in data collection efforts for the Massachusetts pilot study, and project staff experienced in data collection and management. Training in computerized data entry and archiving was given by the Qualidigm representative who designed the software for this study.

A comprehensive training manual was developed specifically for this activity; each assessor was provided with a copy to serve as a reference guide throughout the training course and for the duration of data collection (see Appendix C). During each session throughout the program trainers walked the assessors through each element and demonstrated how to use the manual to clarify issues that come up in the field. Each manual included all field instruments and detailed instructions on how to complete each tool, including sources of information, definitions of key terms, and examples of coding options. Sections on project procedures and resources, maintaining confidentiality, obtaining informed consent, and data management and submittal were also included. Assessors were instructed to use their manuals for trouble-shooting, looking up contact information for key Project staff resources, and for reminders about standard procedures and implementation guidelines.

Following introductory sessions on project activities, the nursing home and post-acute care environments, and roles and responsibilities of nurse assessors, training was comprised of didactic and practical experience in use of all data collection instruments including the Administrative survey, the facility environmental walk-through and observational tool, the medical record review (MRR) tool, and a subset of MDS Version 2.0 items. To assure accuracy and consistency in coding, particular attention was given to providing the assessors with practical experience in coding the latter two instruments and certifying that they could complete them adequately. Following didactic training in the MRR, the assessors worked in small groups to complete reviews using the MRR on up to four nursing home residents medical records (identifying information deleted). Each group was led by a project staff trainer. Case discussion including question and answer periods with the entire group following each MRR session. Certification of skills competency was then completed using a fifth case, followed by one on one remediation with a group leader as necessary.

Two and one-half days of the program were devoted to training in how to conduct resident assessments using a subset of items from MDS Version 2.0 as the assessors were being trained to be the project s gold standard MDS assessors. The didactic portion of the sessions was provided by a clinical nurse specialist with over ten years experience in this area. The training manual included all corresponding guidelines for assessment from CMS s RAI User s Manual. Trainees were instructed to follow the standard assessment processes specified in the RAI User s Manual (Morris et al, 1995) using multiple sources of information (e.g., resident observation, interviews with direct care staff, chart review). Scripted videotaped vignettes were presented to demonstrate interviewing techniques and to provide practice in coding. Trainees were paired for role-playing exercises to practice their interviewing skills. Case presentations and follow-up discussion were used to illustrate assessment techniques and correct coding responses. To certify competency in MDS assessment, each trainee completed a case and met individually with the lead trainer for review.

To enhance and maintain consistency in coding, project staff held weekly one-hour conference calls with the assessors during the course of data collection. Minutes of the calls, which always included a question and answer section, were distributed to all assessors within one week of the call.

3.5 Inter-rater Reliability Among Nurse Researchers

As mentioned in Section 3.4, nurse researchers were assessed for their understanding and ability to correctly complete the MRR and MDS forms prior to leaving the training sessions. In addition to this "certification" process, weekly debriefing teleconferences were held with nurse teams to answer any coding questions. Finally, beginning in January 2002, nurse researchers were required to complete two paired assessments and medical reviews with their partner per facility. Nurses were asked to select cases for inter-rater reliability review at random, once the resident sample at each facility had been selected. Nurses were not to share findings until each of their reviews of both MRR and MDS forms was complete and data entered. Inter-rater review cases were submitted to HRCA along with all other MRR and MDS data submitted from the field. Nurse reviewers were also asked to photocopy portions of the relevant records and submit to HRCA for two main purposes: 1) to provide a context for discussion on subsequent debriefing calls; and 2) to allow spot-checking of results by project nurses.

All nurse reviewers performed some number of the requested inter-rater reliability reviews. Some nurses performed and submitted more inter-rater reviews than others. Agreement statistics for the MDS inter-rater reliability of nurse researchers was very good, and is described in Section 5.2.

In order to understand the degree and nature of nurse assessor item-level consistency, Medical Record Review forms were completed for eight records from five nurse assessor teams representing four states. These reviews were completed as a "spot check" of MRR coding, and were selected based upon lower levels of overall MRR inter-rater agreement between these nurse teams. In general, nurse assessors appeared to pay careful attention to detail and a rationale for their responses could be detected. In many cases where there was disagreement between the project nurse and the assessor(s), it seemed very possible that not all parts of the medical record had been submitted to HRCA for review. Other areas of discrepancy found were attributable to 1) contradictory and/or inconsistent facility documentation; and 2) lack of clear coding instructions. With regard to the latter problem, attempts were made to clarify coding instructions on the weekly debriefing calls, particularly in the area of comprehensive assessments and ADL improvement and decline.

3.6 The Data Collection Process

The overall tasks in management of primary data included: 1) creation of a database to manage all site data obtained during the validation study; 2) data entry of all data collection instruments; 3) processing data submitted by sampled facilities and by nurse reviewers; 4) data cleaning; and 5) identification of sites with complete data for inclusion in the analytical file.

Using Qualidigm's Medquest Clinical Data System Software, field nurses entered MDS supplement and Medical Record Review data onsite, copied the data onto diskettes, and forwarded the diskettes to HRCA, where the data was added to the study database. On numerous occasions, diskettes submitted with MDS supplement and medical record data were found to be empty, corrupt or unreadable and the assessor was asked to submit their backup data. Qualidigm programmers were available by phone to

work with the assessors to resolve problems. When the assessor could not produce a backup copy or Qualidigm was unable to resolve the software problem, assessors were asked to submit paper copies, if available, to HRCA for data processing.

The Administrative Surveys completed by facility staff and copies of the 30 most-recently completed MDS assessments were submitted to HRCA for data entry. HRCA staff also entered the three Environmental/Walk Through Observations and contact sheets for the 30-plus residents from each site.

All site data, including the Administrative Survey, was to be sent by Federal Express to HRCA within the week following the site visit. Although the assessors had been instructed to submit site data in one package, many of them entered their data after the site visit, off-site of the facility; thus, data from a site often arrived in two packages as much as a week apart. The clerk who opened the package checked against the sample roster to verify that each case was complete and the study IDs were correctly recorded on each paper form (contact sheet, consent form when required, facility MDS assessment, and paper or diskette for MDS and medical record review). Diskette data was read and checked for completeness, and errors were corrected prior to being merged into the study database.

Data cleaning programs were written to identify sites with missing data, to ensure that disposition of the case was correctly recorded on the contact sheet, that each MDS assessment was accompanied by the facility MDS completed no more than 90 days earlier. Assessors were contacted by phone or e-mail when IDs could not be matched, or data was incomplete or improperly coded.

HRCA nurses called many facilities repeatedly requesting that Administrative Surveys that had not been completed during the site visit be mailed or faxed directly to HRCA. This process resulted in successfully obtaining all but two of these surveys. Unfortunately, most surveys were returned with at least one missing or questionable response. All problematic items were photocopied and faxed to the facility for correction. When the facility failed to correct one or more items after at least two requests by fax and a follow-up telephone request, these items were coded as refusals.

Ten sites were dropped because of incomplete data:

- Two small hospital-based facilities and one chronic care facility with a small census during the site visit were dropped because the assessors were unable to obtain assessments for a minimum of 20 residents who had been assessed by the facility during the 90-day period prior to the site visit.
- Assessors at one site failed to provide at least 20 complete copies of the most recent MDS assessments by the facility, and attempts to complete the sample were unsuccessful.
- Assessors at one site were denied access to medical records for more than half of the sample; they were told that the records were “locked up.”
- Three sites were lost because data for 15 of the 30 cases could not be recovered from diskette or archive, and paper assessments were not available.
- Two facilities failed to complete the Administrative Survey and were therefore dropped from the analytic sample.

4.0 Methods for Primary Validation of QIs

4.1 Overview

This section of the report presents the methods used in this national study to determine whether a series of MDS-based quality indicators, also referred to as performance measures, are valid measures of the quality of care provided by nursing homes. The analysis is based on a six-state, national sample of nursing facilities (N= 209), and is focused on the relationship between two sets of variables. The first is a series of indicators of nursing home quality based upon aggregated resident data (i.e., the quality indicators). The second are three arrays of measures that relate to service inputs, assessments, and staffing that have been hypothesized as the precursors to good nursing home performance on the quality indicators (i.e., the validation elements). The major premise for these analyses is that if the former quality indicator measures are to be considered meaningful and valid, there should be a significant relationship with the relevant validation measures.

4.2 The Quality Indicators

4.2.1 Description of the QIs Evaluated in this Study

The quality indicators are of two types, "chronic" and "post-acute," and they were derived from one of two sources: they were either in general use in the industry prior to this study, or they were designed by the study team to fill "gaps" in the coverage of the existing indicator set. All post-acute care indicators were created by the study team, as were eight of the chronic measures. The remaining chronic measures were derived from three sources – Ramsey, the Center for Health Systems Research and Analysis (CHSRA), and LTCQ, Inc. The research conducted to select the existing indicators has previously been reported (Abt Associates, Oct-2001).

The largest set of indicators to be tested applies to long-stay residents of nursing facilities. These residents are often referred to as "chronic," with many likely to spend the rest of their lives in a nursing home. These measures do not assess quality at the point of admission, rather, most of them require a minimum exposure period of 90 days in the facility before the indicator comes into play. In fact, for the typical chronic resident, he or she will have been in the facility for more than one year, and in all cases we seek to ensure that to the extent possible, the indicator is an honest reflection of the long-term path of decline of the resident and the intervening care practices of the facility.

There are 38 of these "chronic" quality indicators, divided into prevalence and change-based measures. Twenty-three are prevalence measures, 15 are change-based measures. All but two of the change measures reference declines in status, and these declines occur over a 90-day assessment window (i.e., the scheduled interval between MDS assessments for long-stay, chronic residents). These indicators reference the following dimensions: functional performance; cognition and communication status; mood and behavior; social activities; clinical complications (e.g., incontinence, weight loss, pain, pressure ulcers, infections); falls; use of appliances (i.e., restraints, tubes, catheters); and antipsychotic drug use.

The second type of quality indicator evaluated applies to the short-stay resident population found in skilled nursing facilities. Medicare largely pays for the care for these residents, and a resident of this

type is typically admitted from a hospital and will have a total length of stay of from a few days to a month's duration. These residents are often called subacute or post-acute care (PAC) patients.

Seven PAC QIs were evaluated, referencing the following dimensions: functional performance (i.e., overall ADLs and mobility); delirium; pain; bladder continence; pressure ulcers; and respiratory problems. Two are prevalence-based, five are incidence-based, and all seven reference patient status during the initial two plus weeks of the stay.

4.2.2 The Nursing Home Minimum Data Set

The measures underlying the quality indicators are derived from a facility-mandated, facility-generated, resident assessment tool known as the Minimum Data Set (MDS). CMS first mandated national use of the MDS in 1990, and under this mandate, facility staff are responsible for completing the assessments. And, given this facility assignment feature of the national MDS mandate, it was deemed advisable to reassess the accuracy of these assessments. The quality indicator effort rests on this structure, and for these measures to be usable as inputs into a national quality indicator system, we must be able to "trust" these staff assessments.

MDS reliability reports in the literature from the initial roll out of Version 1 of the MDS in 1990 and Version 2 in 1996 were most positive, although there have been more conflicting assessments reported subsequent to 1996. To further test this issue seemed to be a prudent step in this study. And, while the results are described elsewhere in this report, the bottom line is most encouraging. Facility staff reliability levels remain on par with the earlier reports from the rollouts of Versions 1 and 2 of the MDS. There were no significant inter-state differences in the accuracy of the assessments, and only a handful of facilities appeared to perform poorly (around 5 percent of the total). For a nationally mandated system, these are very positive results, indicating that the U.S. nursing home industry has reacted responsively to this aspect of the federal mandate.

4.2.3 Construction of the Quality Indicators

In constructing the set of quality indicators evaluated here, there has been a concern for possible inter-facility variation in the types of residents admitted and served by the facility; difference in the mix of residents served across facilities raises the possibility that inter-facility comparisons may be biased. To control for this possibility, where deemed necessary, three adjustment strategies have been applied.

1) For all of the indicators a denominator exclusion rule was applied (e.g., residents near death). These residents were not considered in the calculation of the quality indicator.
2) For four of the CHSRA indicators, two sub-versions of the same overall indicator were created for each facility, one applying to high-risk residents, the other to low risk residents. In addition, an overall high/low risk indicator was calculated.
3) For many of the other indicators, including those created by the project team and LTCQ Inc., some type of statistical regression-based covariate adjustment strategy was employed. For many indicators, this involved traditional resident-level covariates, supplemented in many instances by a new type of facility-based adjustment based upon resident characteristics upon admission. This is referred to as the "facility admission profile". QIs constructed using a facility admission profile are designated as such in results table 6.1.

Unadjusted, covariate-adjusted and FAP- and covariate-adjusted quality indicator rates for the set of 45 chronic and PAC QIs were calculated for every facility in the six validation states. Rates were calculated using target quarters corresponding to the time period of primary data collection (i.e., Calendar Quarter 4, 2001 and Calendar Quarter 1, 2002). Adjusted rates were derived from logistic regression models run on a national MDS dataset that consisted of four quarters (excluding the target quarter) of data. Appendix D describes the exact method of QI calculation, and Appendix E contains operational definitions (e.g., numerators, denominators, risk adjustment) of each of the 45 quality indicators.

4.3 Primary Validation of QIs

The primary goal of this full-scale validation study was to determine if the selected set of MDS-based quality indicators reflect the care processes in place in nursing facilities. That is, do the MDS-based QIs measure what they are intended to measure (i.e., validity). Nursing facility quality indicators may be considered valid when they 1) are accurately measured; and 2) reflect a positive relationship between the care reflected by the QI and the care processes and structures in place to achieve those care processes reflected by data collected at a nationally representative sample of nursing facilities. For example, in a facility with a low rate of pressure ulcers (i.e., a "good" facility), we would expect to see care processes in place that are designed to prevent the occurrence of pressure ulcers, or to treat and cure pressure ulcers expediently. The positive relationship between the QI rate and the care processes in place in the facility would allow a determination that the QI in question (in this case, pressure ulcer) was valid.

The accuracy of the measure of quality is of vital importance in any assessment of validity; the analysis of this and related issues is discussed in Section 5. The following section describes the various components of the design of the full-scale validation study and the subsequent development of measures by which QI validity was assessed.

4.3.1 Development of Validation Hypotheses

In facilities with good quality outcomes one should be able to identify care processes and structures that relate to, or could potentially influence, resident outcomes. The project team therefore took a multi-step approach to developing a comprehensive array of observational, survey, and record review tools that could efficiently measure such processes and structures. This process was previously described in Section 3.3.

As described, the data collection tools and subsequent validation analyses were based upon a series of hypotheses regarding the relationship of "good" care or best practices in nursing facilities to good quality outcomes in specific care dimensions. One example is provided here to further articulate this process.

The expert clinical panel that dealt with the "Pressure ulcer" quality dimension developed a series of hypotheses related to the ability of the facility to minimize the incidence of pressure ulcers among their residents or to manage the patient with a wound admitted from other settings of care. The expert panel reviewed clinical practice guidelines regarding pressure ulcers from both the Agency for Health Research and Quality (formerly the Agency for Health Care Policy and Research) and the American Medical Directors Association in proposing these hypotheses. The pressure ulcer hypotheses include (but are not limited to):

Hypothesis 1: Facilities that have the following in place will have fewer new pressure ulcers arise among their patient population:

- a standardized assessment protocol for identifying the patients at risk,
- policies and procedures to specifically address the individual's risk factors, and
- explicit programs for implementation and monitoring of individualized prevention interventions.

Hypothesis 2: Facilities that have surveillance mechanisms to identify early signs of tissue injury will have fewer new pressure ulcers arise among their patient population.

Hypothesis 3: For patients with pressure ulcers, attention to support surfaces, positioning protocols and padding will result in fewer new pressure ulcers among these patients.

The project team utilized these hypotheses both to create data collection items on the medical record review, environmental observation and administrative questionnaire and to form validation constructs upon data analysis. In implementing data collection for this area of care quality, medical records were reviewed to determine if sampled facilities used screening tools or other assessments (i.e., Norton or Braden scales), research nurses observed during the environmental tour if positioning devices were in use, and Directors of Nursing were asked about facility pressure ulcer policies, quality improvement activities, and educational efforts regarding the prevention of pressure ulcers.

4.3.2 Process of Developing Final Validation Scales

In addition to guiding the content of data collection instruments, the hypotheses for quality dimensions were also used to construct validation scales or "constructs" by which to assess facility quality in the analysis of validation data.

In the full-scale validation study, emphasis was placed upon 21 key dimensions of quality: cognitive impairment, communication, delirium, depression/mood, behavior problems, ADL improvement, ADL decline, mobility/walking, falls, antipsychotic drugs, pain, physical restraints, feeding tubes, undernutrition, indwelling urinary catheter, bladder incontinence, bowel incontinence, infections, pressure ulcers, burns/abrasions/skin tears, and little or no involvement in activities. In addition to these dimensions, data regarding facility paradigms such as a preventative, enhancement-oriented approach, good and comprehensive care planning and assessment processes, and access to consultants in and outside the facility were gathered during the site visits.

During the analysis phase of the validation study, a series of activities occurred:

- Re-examination of validation scales/constructs used in the pilot study;
- Creation of new validation constructs for all 21 quality dimensions; and
- Examination of the relationship of individual data collection instruments to quality dimensions/quality indicators.

Each validation scale used in the pilot analysis was re-examined to determine the expected strength of the relationship with the quality indicator, and with other quality indicators if warranted. In addition, data collection instruments were reviewed in order to identify validation elements for specific quality

indicators beyond the eight primary dimensions targeted by the pilot study, and to construct validation scales that may reflect performance in multiple domains of quality.

Methods

Project members reviewed all available data collection instruments, and examined the frequency distribution of each data element of interest. They suggested individual items, combination of elements or summary scales based on content and the distribution of responses. That is, items with no variation (e.g., all facility responses on a given item were "yes") were not used because they would not discriminate between good and poor performers.

Each proposed validation scale was discussed by the project team. One hundred seventy- four-validation scales in all were created or re-examined from the pilot study (see Appendix F). The clinical validity of each scale was reviewed, as was the frequency distribution of the scale to ensure that it demonstrated sufficient variation. Based on conference discussion, some scales were modified and others deleted. If similar constructs were addressed by more than one scale, preference was given to scales with better potential for applicability to multiple domains and to those with variation in the distribution of responses. If merited, judgments were made as to whether the hypothesized relationship between the validation construct (or scale) and the QI was expected to be moderate (Level 1) or weak (Level 2).

Finally, relationships between individual data collection instruments (either in entirety or by individual data item) and QIs were examined to determine the strength of these relationships. This evaluation revealed that many of the medical record review items, for example, bore a strong relationship to individual quality indicators, absent any additional data elements or a priori construct. Again, these data items were collected because it was hypothesized that the processes they measured (e.g., care planning, comprehensive assessment) were related to quality, so these positive relationships were not unexpected.

4.3.3 Final Validation Constructs

The final validation elements utilized to determine degree of quality indicator validity were categorized as follows:

- ***Preventive*** strategies represent the class of actions that "good" facilities choose to follow in an attempt to minimize the emergence of problems; these strategies are anticipatory in character. Data elements categorized into the preventive construct include staff training, higher staff resource levels, and facility efforts at continuous quality improvement (CQI).

- ***Responsive*** strategies represent actions that facilities are likely to use as they recognize that residents have ongoing or emerging problems in different quality areas. They represent a service response "audit trail," and as such confirm that staff have recognized the problem. Externally, these facilities will be observed to have higher QI scores, but the medical record will reflect a recognition that action must be taken in response to identified resident problems. Examples of data elements gathered on-site that are categorized as responsive are the documentation of comprehensive assessments (other than the MDS), documentation of changes in resident status, and referrals to specialists from inside and outside of the facility (e.g., physicians).

In summary, preventive strategies work to reduce the prevalence or incidence of quality problems measured by the QIs. On the other hand, responsive strategies reflect the fact that quality problems may have emerged in the resident population and as such reflect a "failure" of the facility to prevent the problem (or failure to achieve expected improvement outcomes). Consequently, responsive strategies are associated with an increased prevalence of problems (i.e. quality indicators).

While the constructs created from the various sources of data were conceptualized as falling into one class or another, clinically and administratively relevant data elements thought to be related to particular QIs might have been able to be classified as either preventive or responsive. Thus, our final classification of the validation elements was done based both upon how they related to one another as well as how they related to the QIs. While seeming to represent a "circular" logic (i.e. using one construct to validate another and then to apply the same logic in the other direction), this is a process that characterizes most efforts at construct validation. Thus, the validation elements and constructs were examined for directionality relative to selected QIs and then the QIs were each formally tested against the battery of constructs (classified as preventive or responsive) to determine whether facility records, care processes and structures related to the QIs in the expected direction.

For each of the constructs or individual data elements categorized as preventive or responsive, the relationship between it and the full array of quality indicators under study was reviewed. To be found acceptable, the construct had to have a consistent relationship across multiple quality indicators. For example, to be classified as preventive, a data element (e.g., a CQI monitoring protocol) had to be related to several quality indicators. We required that there be a clear directional relationship between the construct and the quality indictors. Specifically, preventive elements had to always show a positive relationship to lower (less problematic) QI rates, while responsive elements had to demonstrate a positive relationship to higher (more problematic) QI rates. In other words, the correlation between preventive elements and quality indicators had to be negative, and the correlation between responsive elements and quality indicator scores had to be positive to be considered clearly directional.

In evaluating the validity of the quality indicators, several summary measures were created and reviewed:

- a count of the number of significant preventive or responsive validation elements for the quality indicator, with the greater the count, the greater the confidence in the relationship;
- a measure of the pooled association of the list of significant validation elements with the quality indicator. The latter is derived from a regression equation, and in this case represented by the multiple correlation coefficient. This is a multivariate-derived value that resembles a standard bivariate correlation[9]. In reviewing these values, we settled on a combination of two factors in assigning each of the candidate quality indicators to one of three "valid" categories: Top, Mid, and Not Validated; and
- the underlying reliability of the MDS item and resulting QI.

To understand how these preventive and responsive factors were applied in establishing the validity of a QI, we provide examples of how these elements individually relate to two of the chronic quality indicators, "Pressure ulcer prevalence" (high & low risk) and "ADL worsening". Both indicators are assigned to the Level I, Top Validity category, and both achieved this status on the basis of the

[9] Note: this value can be squared to get the classic R2 estimate of explained variance.

preventive elements alone. For the Pressure ulcer indicator, there was also a substantial array of individual responsive relationships, while for the other, ADL worsening, there was only one item of this type.

The ***Pressure ulcer*** indicator quantifies the proportion of at-risk residents in a facility that have a pressure ulcer (i.e., bed sore, decubitus ulcer, pressure sore) of severity ranging from one persistent area of redness that does not disappear when pressure is relieved to one or more open wounds where the full thickness of skin and subcutaneous tissue is lost and underlying bone or muscle is exposed.

There are a large number of clinical and functional risk factors for pressure ulcers (e.g., poor nutrition, incontinence, diabetes, immobility); thus, a number of preventive activities and responsive factors were evaluated. Preventive activities, in general, relate to the handling of at-risk residents and treatment of conditions that contribute to or mitigate pressure ulcer risk. Responsive activities, in general, define actions that a facility's caregivers take to document, communicate and attempt to ameliorate pressure ulcers once present.

Preventive activities for pressure ulcer prevalence included the screening, assessment, and treatment for conditions placing residents at risk for pressure ulcers. Thus, the following individual data elements or constructs were found to be associated with lower pressure ulcer prevalence:

- More frequent scheduling of assessments for suspicious skin areas.
- Weekly routine assessment using a standard protocol for delirium, that would - if present - keep residents bed-bound.
- Observations on the environmental assessment of residents walking or otherwise out of bed.
- Observations on the environmental assessment of caregivers providing assistance to residents with nutritional needs.
- A constructed scale expressing the extent to which a facility manages clinical, psychosocial, and nutritional complications across domains in a manner consistent with high quality care (expressed as a single factor score).

Staffing factors provide additional (albeit indirect) evidence of preventive activities. For example, staffing items related to pressure ulcer prevalence were 1) the absence of facility management change; and 2) the extent that a facility did *not* rely upon floats or contract staff.

Responsive activities for pressure ulcer prevalence include policies, procedures or actions taken by caregivers in response to existing or newly detected pressure ulcers. Identified activities include:

- Comprehensive assessment (other than the MDS) of pressure ulcers documented in the medical record.
- Assessment of pressure ulcers by a physician.
- Clear documentation in the medical record that the resident has a problem in this area or that the resident's condition has changed relative to pressure ulcers.
- Where change was noted in the medical record, there is documentation that this change 1) was evaluated within 72 hours, 2) resulted in a notification to physician or therapist, 3) resulted in a referral to a consultant, and/or 4) resulted in a change in the care plan.

An additional theme related to pressure ulcers was a constructed measure of the extent to which the medical record and care plan agree that pressure ulcers are a problem. This level of agreement signals facilities with a well-integrated system for problem recognition and treatment implementation.
For ***ADL worsening***, there were 17 significant preventive elements and one significant responsive element. From this set of preventive elements, three primary themes emerge: attention to the resident as an individual, an engaging and safe environment, and good continence care. Further explanation of these themes and related data elements follows.

- Maintaining ADL gains is related to a concern with what the resident is thinking and who he or she may be as a person, as seen in areas related to cognition, behavior, and pain. Better outcomes (i.e. facilities have lower rates of ADL worsening) are observed when there are: 1) CQI monitoring protocols in place for behavioral function and communication; 2) weekly routine screening of communication and pain, using standard protocols; and 3) rooms that are personalized with furniture, photos, and other things from the resident's past.

- Maintenance of ADLs is also related to an environment in which the resident is up and out of bed and engaged in activities. Better outcomes are related to a series of things that were observed by the research nurse about the facility, including: 1) residents being up and about; 2) residents seen to be walking or independently moving about the facility with or without assistive devices; and 3) indications that a variety of activities are available for residents with different capabilities. Related data elements observed during inspection of the facility environment were that public and common areas were well lighted and resident safety had been considered.

Finally, there was a link to facility efforts aimed at good continence care. Preventive elements relating to this theme include: 1) a scale that counted up to 15 "good" incontinence management items; 2) a scale that focused on care practices relevant to promoting improved levels of continence; 3) a scale that looked specifically at ADL training approaches that were targeted to helping residents maintain continence patterns; and 4) a CQI monitoring protocol in place for bladder incontinence.

5.0 Methods for Evaluating Reliability and Measurement Bias

Prior to conducting the analyses to validate the meaningfulness of the quality indicators, it was crucial to first establish whether the data submitted by participating facilities were reliable and without substantial measurement bias. Since the QIs are based upon MDS data submitted on all residents and admissions from all facilities, if a facility's data are consistently unreliable or biased in a particular manner, that facility's data would increase the noise, or error, in the data being used to test the validity of the QIs. To the extent that this occurs, our test of the validity of the QIs will be compromised. Consequently, one of the principle reasons for conducting reliability and measurement bias analyses was to consider dropping facilities from the pool of facilities included in the analyses. The methods used to test the reliability and measurement bias in the data are described in the sections below.

5.1 Testing for Inter-rater Reliability

The data collection effort in each facility had the research nurses gather over 100 different MDS data elements independently about each sampled nursing home patient, both new admissions and long-stay residents. These MDS assessments were done as part of the research nurse's examination of sampled patients' medical records as well as their observation of, or conversation with, the patient. In comparing the MDS assessment elements recorded in the facility MDS with those recorded by the study research nurses, we had to ensure that the research nurses were reliable one with the other. Thus, the first step in the testing for inter-rater reliability was testing the inter-rater reliability among the research nurses. These nurses under went a central training by HRCA nurses, all of whom have extensive MDS experience, and have worked with CMS in the design and refinement of the MDS since its inception in 1990.

In most participating study facilities, pairs of research nurses worked together to split the work and to ensure efficient conduct of the entire data collection protocol. One feature of that was to have the two research nurses conduct an inter-rater reliability test on several residents in many of the study facilities (see Section 3.5). While there were not enough residents assessed by the same pair of raters to permit inter-rater reliability assessments for each research nurse, it was possible to pool the paired reliability assessments done among the research nurses. In this way, we established the general inter-rater reliability of the research nurses. To the extent that they are found to be reliable, one can assume that comparisons to any one research nurse are generalizeable to all others. Furthermore, since the goal was to test not only the degree of inter-rater reliability in the study facilities, but also the extent to which there is measurement bias, it is important to know that the research nurses can be thought of as the "gold standard" against which the measurement performance of the facility nurses can be compared.

The approach used to test inter-rater reliability is the Kappa statistic, or the weighted Kappa for ordinal measures such as ADL performance, etc (Cohen, 1960). This statistic essentially compares the two sets of raters who have each observed and assessed the same patient independently. However, rather than merely calculate the percentage of cases on which they agree, the Kappa statistic corrects for "chance" agreement, where "chance" is a function of the prevalence of the condition being assessed. It is possible that two raters could agree 98 percent of the time that a resident had episodes

of disorganized speech. However, it might be the case that virtually no residents were rated by either rater as having episodes of disorganized speech and that they never agreed when one thought that the condition was present. In this instance, in spite of the fact that the level of agreement would be very high, the Kappa would be very low. Depending upon the importance of the assessment construct, or item, having a low Kappa in the face of very high agreement and high prevalence could be very problematic or a trivial concern. For this reason, we will generally present the percentage agreement as well as the Kappa, or weighted Kappa. The weighted and unweighted Kappas are identical for dichotomous (binary) measures such as all the Quality Indicators (presence or absence); however, the ordinal measures like ADL or cognitive decision-making are more appropriately assessed with the weighted Kappa.

By convention, a Kappa statistic that is .70 or higher is excellent whereas a Kappa statistic that is less than .4 is considered unacceptable. Levels in-between are acceptable. These standards were applied for both the individual MDS data elements as well as the composite, dichotomous quality indicators.

The total number of pairs of observations for the inter-rater reliability analyses is nearly 4,000. Obviously, in view of the numbers of observations, any estimate of the overall degree of inter-rater reliability for the sample of participating facilities in this study will be very stable. For the most part, the number of pairs of observations per facility is between 25 and 30. This number of observations yields a fairly stable estimate of inter-rater reliability to characterize the facility, given that the observations are representative of the residents and nurse raters in the facility and conditional on the relative prevalence and distribution of the condition (e.g., dementia, pain) in the facility. This means that a Kappa statistic characterizing the reliability of all raters in a given facility is likely to reflect the stability and commonality of assessment perspectives among individuals in the home.

In some instances, particularly for calculating the QIs at the level of the individual resident, restrictions on the residents to which a QI applies ends up reducing the number of paired comparisons within a facility as the basis for calculating a facility-specific Kappa. In order to insure that a minimum number of observations are included in the calculation of Kappa, we set the threshold at five. The confidence intervals around an estimate of the Kappa is a function of the absolute percentage agreement, the prevalence, or variance, of the condition as well as the number of pairs being compared. Holding constant the prevalence and agreement rate, the size of the confidence interval is clearly related to the number of observations. For a facility with 30 paired observations, the approximate 95 percent confidence interval is +/- .25; that interval becomes +/- .65 when there are only five observations. While it is obvious that this means that the confidence interval around the Kappa estimate for a given facility may be quite large, we didn't want to lose significant numbers of facilities to the inter-rater reliability analysis by requiring the number of pairs to be much higher.

5.2 The Reliability of "Gold Standard" Research Nurses

In all, a total of 119 pairs of resident assessments were compared across the 26 research nurses included in the validation study. Each research nurse rated from one to 17 of the paired assessments. The results of the analyses are presented below in Table 5.1.

Table 5.1
Summary Inter-Rater Reliability Statistics of MDS items for Research Nurses

MDS Item		Percent Agreement	Kappa	Weighted Kappa*
A10A	Living Will	87.16	0.61	0.61
A10B	Do Not Resuscitate	91.45	0.83	0.83
A10C	Do Not Hospitalize	97.22	0.39	0.39
A10F	Feeding Restrictions	97.25	0.89	0.89
A10G	Medication Restrictions	97.22	0.83	0.83
A10H	Other Treatment Restrictions	92.59	0.69	0.69
A10I	Advanced Directives: None Above	96.33	0.93	0.93
B2A	Short-term Memory	88.24	0.63	0.63
B4	Cog Skill for Daily Decision Making	97.29	0.85	0.89
B5A	Less Alert, Easily Distracted	97.88	0.85	0.79
B5B	Periods of Altered Perception	97.69	0.80	0.75
B5C	Episodes of Disorganized Speech	97.69	0.79	0.72
B5D	Periods of Motor Restlessness	96.22	0.67	0.66
B5E	Periods of Lethargy	98.11	0.80	0.78
B5F	Mental Function Varies over the Day	96.64	0.78	0.71
C4	Making Self Understood	95.89	0.73	0.82
C6	Ability to Understand Others	96.08	0.76	0.80
E1A	Patient Made Negative Statements	98.32	0.89	0.89
E1C	Repetitive Verbalizations	98.11	0.65	0.71
E1D	Persistent Anger with Self/Others	98.95	0.84	0.86
E1E	Self Deprecation	97.48	0.56	0.56
E1F	Express Unrealistic Fears	96.64	0.76	0.76
E1G	Recurring State - Something Terrible	99.16	0.80	0.80
E1H	Repetitive Health Complaints	94.12	0.73	0.73
E1I	Repetitive Anxious Complaints	97.69	0.74	0.73
E1L	Sad, Pained Facial Expression	95.38	0.68	0.71
E1M	Crying, Tearfulness	98.32	0.74	0.78
E1N	Repetitive Physical Movements	97.27	0.77	0.86
E2	Mood Persistence	94.49	0.73	0.81
E4A.A	Frequent Wandering	98.79	0.85	0.85
E4B.A	Frequent Verbally Abusive	100.00	1.00	1.00
E4C.A	Frequent Physically Abusive	98.69	0.76	0.74
E4D.A	Frequent Socially Inappropriate Behavior	99.35	0.75	0.87
G1AA	Bed Mobility Self-Perform	96.02	0.72	0.86
G1BA	Transfer Self-Perform	97.80	0.71	0.92
G1CA	Walk in Room Self-Perform	97.01	0.72	0.91
G1DA	Walk in Corridor Self-Perform	95.23	0.74	0.86
G1EA	Loco on Unit-Perform	94.81	0.71	0.85
G1FA	Loco off Unit Self-Perform	96.28	0.74	0.89
G1GA	Dressing Self-Perform	96.59	0.69	0.85
G1HA	Eating Self-Perform	96.96	0.84	0.88
G1IA	Toilet Use Self-Perform	97.59	0.76	0.91
G1JA	Personal Hygiene Self-Perform	96.96	0.70	0.89
G8A	Res-increased Independence in some ADLs	93.52	0.74	0.74
G8B	Staff-increased Independence in some ADLs	89.19	0.73	0.73

Table 5.1
Summary Inter-Rater Reliability Statistics of MDS items for Research Nurses

	MDS Item	Percent Agreement	Kappa	Weighted Kappa*
G8C	R Able to Perform Tasks Slowly	90.00	0.47	0.47
G8D	Major Diff ADLs Morning vs. Evening	95.37	0.26	0.26
G8E	ADL Rehab Potent: None Above	86.96	0.72	0.72
H1A	Bowel Continence	94.96	0.77	0.88
H1B	Bladder Continence	95.70	0.78	0.88
H3D	Indwelling Catheter	97.22	0.79	0.79
H3E	Intermittent Catheter	99.08	0.80	0.80
H3F	Didn't Use Toilet Room	91.74	0.53	0.53
H3G	Pads/Briefs Used	89.47	0.78	0.78
H3I	Ostomy	99.08	0.80	0.80
H3J	Appliance and Programs: None	90.99	0.81	0.81
I1FF	Manic Depressive	100.00	1.00	1.00
I1GG	Schizophrenia	99.07	0.90	0.90
I1RR	Diseases: None of the Above	96.64	0.78	0.78
I1X	Paraplegia	97.22	0.39	0.39
I2E	Pneumonia	99.08	0.85	0.85
I2F	Respiratory Infection	98.15	0.89	0.89
I2G	Septicemia	100.00	1.00	1.00
I2J	Urinary Tract Infection	96.36	0.88	0.88
I2L	Wound Infection	99.07	0.80	0.80
I2M	Infection: None of the Above	93.97	0.85	0.85
J1B	Unable to Lie Flat	95.45	0.59	0.59
J1H	Fever	99.07	0.88	0.88
J1I	Hallucinations	100.00	1.00	1.00
J1L	Shortness of Breath	91.82	0.71	0.71
J1P	None of the Above	93.04	0.78	0.78
J2A	Pain Frequency	92.95	0.72	0.78
J2B	Pain Intensity	98.18	0.73	0.82
J4A	Fell Past 30 Days	93.75	0.00	0.00
K3A	Weight Loss	97.46	0.83	0.83
K5B	Feeding Tube	99.08	0.92	0.92
K5C	Mechanically Altered Diet	90.99	0.82	0.82
K5I	Nutritional Approach: None Above	92.11	0.84	0.84
M2A	Pressure Ulcers	98.73	0.73	0.83
M4F	Skin Tears	95.37	0.76	0.76
M4H	Other Skin Problems: None of the Above	94.92	0.72	0.72
N2	Average Time Involved in Activities	95.34	0.57	0.65
O4A	Days Received: Antipsychotics	97.32	0.91	0.92
P1AO	Spec Program: Hospice	99.07	0.66	0.66
P1AS	Spec Program: None of the Above	99.16	0.66	0.66
P4C	Restraints: Trunk Restraint	98.09	0.66	0.72
P4E	Restraints: Chair Prevents Rising	97.01	0.74	0.80
Q1C	Discharge Planned Within 3 Months	95.06	0.76	0.66

Notes:
* weight = 1-[(i-j)^2/(g-1)^2] where i, j are row and column number, and g the number of groups
weighted Kappa inflated with the function sbicc = (2*kw) /(2*kw + (1-kw)) where kw is the weighted Kappa

As can be seen, the percentage agreement and the weighted and unweighted Kappa statistics are high for most MDS items. Only three elements (shaded in gray) have a Kappa that is below .4 (the accepted minimum, particularly for highly skewed variables), and these are highly prevalent and not incorporated into any of the quality measures. The average weighted Kappa for all 87 items is .78, well within the excellent range. Thus, it is clear that these research nurses were well trained and behaved in a similar manner, meaning that all inter-rater reliability performance comparisons between the research and facility nurses can be compared.

5.3 Estimating the Extent of Systematic Measurement Bias

While inter-rater reliability provides evidence of the degree of agreement, correcting for chance, between "gold standard" nurse assessors and facility nurses, even an acceptable Kappa still leaves room for the possibility that all disagreements between raters are in the same direction. For example, the Kappa between the research and facility nurses for one of the measures characterizing the presence of behavioral problems might be .6. The Kappa statistic provides no indication of the "directionality" of the disagreements, but it may well be that facility nurses "normalize" such manifestations of behavioral disturbances and so are less likely to record them as present than are the research nurses. In this way, there is a measurement bias toward under-reporting, or minimizing the presence of selected kinds of clinical problems.

The rationale for exploring the presence of "measurement bias" relates to one of our concerns about comparing nursing facilities across the country using the QIs. Examinations of national data on the prevalence of conditions like pain have found that there is substantially less pain reported among residents in some states than in others, in spite of the fact that the clinical characteristics of nursing home residents in those different facilities is quite similar. Similarly, anecdotal evidence suggested that some facilities focused more aggressively on the identification of some clinical problems such as behavioral problems, distressed mood and pain than did others. The relevance of this suggestion for the development and dissemination of QIs is that facilities that more aggressively identify clinical problems in the MDS assessment will be ranked as performing worse with respect to those QI areas. Since our "gold standard" research nurses were uniformly trained and were found to be reliable, one to the other, they provided an ideal opportunity to see how facility nurses in our participating facilities assessed some of these subjective states relative to a common standard – the research nurses. To address this issue, we created a statistic that estimates the extent to which there is a consistent direction to the disagreement between raters. There might be very limited or considerable disagreement between two raters, but as long as those disagreements are not consistently in one direction or another, there is no bias.

We are interested in comparing the results from our 'gold raters' to the facility raters for each of the QIs. Our trained raters are considered the 'gold standard' because *a priori* there is no reason to believe that they will over or under report any of the QIs for some facilities (i.e., there should be a consistency across facilities). There are many statistical methods that could be used for comparing the raters. For a review (see Banjeree, et al, 1999). The most common statistic for assessing agreement between two raters for binary random variables is Cohen's Kappa (Cohen, 1960). A feature of the Kappa statistic is that it adjusts for the probability of agreement by chance.

For the measurement bias analyses we are interested not only in whether the raters agree, but also in whether disagreement is systematic within facilities. That is, our interest is in determining whether or

not facilities tend to over or under report each indicator. We measure the chance-corrected measure of disagreement using the following statistic, which we will refer to as Gamma. We will index facility by *i* and patient by *j*. Let G_{ij} be the value of the indicator from the gold rater for facility *i* and patient *j*. Similarly, F_{ij} is the indicator from the facility rater. There are two types of errors that can be made (false positive and false negative). In the spirit of Kappa, we penalize each error for the probability of disagreement by chance. This leads to chance-corrected directional Kappa-like statistic, Gamma,

$$\gamma_{ij} = P(F_{ij}=1 \mid G_{ij}=0)/P(F_{ij}=1) - P(F_{ij}=0 \mid G_{ij}=1)/P(F_{ij}=0).$$

That is, γ_{ij} is the difference in false positive and false negative rates, except that each rate is 'adjusted' for the probability of disagreement be chance, e.g., if the prevalence of the indicator is low, then a false positive is considered a more serious mistake. Positive values of Gamma indicate that the facility tends to over report the indicator; Gamma equal 0 indicates that the facility does not under or over report the indicator, on average; negative values correspond to under reporting.

We conducted a simulation study to determine a 'rule of thumb' for classifying facilities based on Gamma. Data were simulated from 10,000 facilities, where each facility's data were generated under one of the following five scenarios: 1) large negative disagreement (facility raters under report the QI); 2) small negative disagreement; 3) no direction to the disagreement; 4) small positive disagreement; and 5) large positive disagreement. Based on the simulations, we classify Gamma in an analogous way to Landis and Koch's (1977) classification of Kappa as follows: 1) Gamma<-0.6 is large negative bias; 2) -0.6<Gamma<-0.2 is moderate negative bias; 3) -0.2<Gamma<0.2 is little to no bias; 4) 0.2<Gamma<0.6 is moderate positive bias; and 5) Gamma>0.6 is large positive bias.

We generated the Gamma statistic per facility for all facilities with at least five paired inter-rater reliability observations. The basic data per facility generated and included in Appendix G is the prevalence for the "gold standard" and for the facility raters, the false positive and the false negative rate (assuming that the research nurse is the "gold standard), the facility Kappa and the resulting facility Gamma statistic. Since we anticipated inter-state differences in the directionality of the measurement bias, we also chose to report the distribution of the Gamma statistic separately by state for each QI. At the level of the facility we plotted the distribution of the Gamma statistic on each QI as a histogram to provide an indication of the directionality of the participating facilities' assessments. Finally, we cross-tabulated the frequency of QIs being in the high negative or in the high positive across all facilities so as to provide an overall assessment of whether, relative to the research nurse assessors, participating facilities were under or over-reporting problems.

5.4 Analyzing the Relationship Between Measurement Bias and the QI

We conducted descriptive graphical analyses as well as multiple regression analyses in order to determine if the Gamma statistic moderates the relationship between the facility Quality Indicator measurement and the facility admission profile (FAP). The QI is a measure indicating the proportion of residents in the facility with a given condition, based upon the most recently available facility-wide MDS data. The QI is based upon the prevalence "snap shot" population of nursing home residents. The FAP reflects the proportion of individuals admitted to the facility in the 12 months prior to the measurement of the QI with the condition that would otherwise trigger them to meet the QI condition.

The graphical analyses were done by creating a scatterplot of the relationship between the QI and the FAP, with different colors indicating those facilities with a substantial negative vs. a substantial positive Gamma statistic vs. those with a Gamma statistic around zero (0). The regression analysis was done regressing the FAP on the QI, controlling for two indicator variables based upon the Gamma – one suggesting a large positive Gamma and the other suggesting a large negative Gamma, with the Gammas around zero serving as the referent group. In conducting these analyses, we focused both on the relative strength of the Gamma associations as well as the extent to which the relationship between the FAP and the QI changes with the introduction of the Gamma in the regression model. These findings are described in Section 7.0.

6.0 Results

As has been stated previously, the reliability and validity of quality indicators is of utmost importance in any deliberations regarding QI utility. This section presents findings on MDS and quality indicator reliability, and on the presence (or absence) of systematic measurement bias (or "ascertainment bias") in this set of evaluated quality indicators. In addition, we report on the degree of validity of each of the 45 tested QIs.

6.1 Reliability/Ascertainment Bias Findings

We undertook the reliability and ascertainment bias analyses for several different purposes. First, in order to provide the best test of the validity of the quality indicators, we wanted to determine whether poor MDS data quality might adversely affect our ability to detect a relationship between the validation elements and the various quality indicators. If we were to find that the overall reliability of the MDS data was poor, the strength of any validation effort would be seriously questioned. The reason for gathering sufficient information to determine the reliability of the MDS data in each participating facility was to allow us to exclude facilities that revealed systematic data reliability problems across a broad range of MDS data elements used to create QIs.

We also undertook these analyses to test the possible influence of systematic measurement bias on the reliability and validity of the QIs. Based upon comparisons of the prevalence of selected clinical care problems from state to state, we surmised that assessors in some areas of the country were more or less likely to assess residents as having some care problems that are used in the construction of QIs. Thus, treating our research nurse assessor as the "gold standard", we sought to understand the extent to which systematic bias (facility assessors tending to miss problems when research assessors found them or *vice versa*) existed for each QI in each facility and ultimately how it related to the QI.

6.1.1 MDS Reliability

Reliability was evaluated in several ways. Research nurse MDS assessments were compared to facility-generated MDS assessments to generate the following statistics: 1) percent agreement between "gold" standard nurses and facility nurses; 2) MDS item-level Kappas; and 3) Kappas for a subset of the QI where these could be established (i.e., for prevalence QIs only).

Table 6.1 displays reliability and distributional statistics for each of the quality indicators for the 209 facilities in the national study sample. Reliability was assessed using the weighted Kappa statistic, with a value of .40 or higher being considered indicative of inter-assessor agreement, while a value of .75 or higher is indicative of superior inter-assessor reliability. In this case the weighted Kappas reflect the cross-sectional reliability of the MDS items that comprise the numerator of the quality indicator (e.g., the numerator for the "Falls" QI is MDS item J4a). Using this standard, only one of the MDS items for a QI numerator falls below the .40 threshold (MDS item N2, which makes up the "Little to no activity" QI). Thirty-one of the quality indicators are based on MDS items with an average weighted Kappa of .70 or higher.

Table 6.1 also displays the mean rates of the quality indicators across the 209 sampled facilities. As seen here, only two quality indicators have very low prevalence (i.e. < five percent). The rate of the chronic care "New insertion of indwelling catheter" indicator is two percent, and the rate of the post-

acute care "Failure to improve and manage delirium" indicator is three percent across the sampled facilities. Five of these QIs have very high prevalence (i.e., > 60 percent). The rate of "Bladder and bowel incontinence – high and low risk" is 62 percent, the rate of "Bladder and bowel continence – high risk" is 93 percent. Similarly, the chronic care "Improvement in walking" indicator is 82 percent. Two post-acute care indicators, "Failure to improve during the early post-acute period" and "Failure to improve or prevent respiratory problems" have rates of 63 and 92 percent, respectively. The rate of occurrence of various QIs is another criterion that should be taken into consideration when evaluating the utility of various QIs, as extreme skews in the rates of occurrence may indicate QI instability, as well as poor utility in detecting inter-facility variation.

Table 6.1

QI Rates and Weighted Kappas

Quality Indicator	QI Proportional Rate – The Average Across Facilities	Standard Deviation of the QI Rate	The Rate in the Facility with the Lowest Proportional Problem	The Rate in the Facility with the Highest Proportional Problem	Average Weighted Kappa for MDS Items Composing the QI [1]
Chronic Prevalence					
++Behavior symptoms (high&low risk)BEH1	.20	.10	.00	.68	.71
++Behavior symptoms (high risk) BEH2	.23	.11	.00	.69	.71
++Behavior symptoms (low risk) BEH3	.07	.05	.00	.23	.71
Little or no activity SOC2	.12	.12	.00	.77	.28
Prevalence of indwelling catheter CAT2	.07	.05	.00	.32	.71
++Bladder/bowel incontinence (high&low risk) CNT1	.62	.13	.14	.89	.88
++Bladder/bowel incontinence (high risk) CNT5	.93	.05	.76	.99	.88
++Bladder/bowel incontinence (low risk) CNT6	.49	.13	.12	.83	.88

Table 6.1

QI Rates and Weighted Kappas

Quality Indicator	QI Proportional Rate – The Average Across Facilities	Standard Deviation of the QI Rate	The Rate in the Facility with the Lowest Proportional Problem	The Rate in the Facility with the Highest Proportional Problem	Average Weighted Kappa for MDS Items Composing the QI [1]
Urinary tract infections CNT4	.08	.05	.00	.31	.53
Falls FAL1	.08	.04	.00	.24	.52
++Infections (pilot) INFX	**.17**	**.08**	**.00**	**.43**	**.50**
++Feeding Tubes NUT1	.08	.05	.00	.27	.80
++Low Body Mass Index BMIX	.12	.05	.00	.31	.85
++Weight loss (pilot) WGT1	**.08**	**.04**	**.00**	**.26**	**.42**
++Inadequate Pain Management (pilot) PAIX	**.11**	**.08**	**.00**	**.48**	**.73**
++Pressure ulcers (high&low risk) (pilot) PRU1	**.09**	**.05**	**.00**	**.27**	**.74**
++Pressure ulcers (high risk) PRU2	*	*	*	*	*
++Pressure ulcers (low risk) PRU3	*	*	*	*	*
++Burns, skin tears or cuts BURX	.05	.04	.00	.19	.46
Restraints used daily (pilot) RES1	**.07**	**.09**	**.00**	**.49**	**.56**
++Antipsychotic use (high&low risk) (pilot) DRG1	**.21**	**.08**	**.02**	**.43**	**.89**
++Antipsychotic use (high risk) DRG2	.43	.11	.26	.61	.89
++Antipsychotic use (low risk) DRG3	.17	.07	.02	.40	.89

Table 6.1

QI Rates and Weighted Kappas

Quality Indicator	QI Proportional Rate – The Average Across Facilities	Standard Deviation of the QI Rate	The Rate in the Facility with the Lowest Proportional Problem	The Rate in the Facility with the Highest Proportional Problem	Average Weighted Kappa for MDS Items Composing the QI [1]
Chronic Incidence					
Late-loss ADL worsening (pilot) ADL1	**.16**	**.09**	**.00**	**.44**	**.84**
ADL worsening ADL2	.08	.07	.00	.33	.83
ADL improvement ADL3	.25	.09	.08	.48	.83
++Locomotion worsening MOB1	.14	.07	.01	.40	.82
++Improvement in walking WALX	.82	.08	.61	.99	.84
++Cognition worsening COG1	.12	.07	.00	.43	.76
++Worsening communication COM1	.11	.07	.00	.31	.83
++Delirium DELX	.09	.06	.00	.29	.61
++Worsening behavior BEH4	.07	.05	.00	.24	.72
++Depressed anxious mood worsening MOD3	.15	.07	.00	.37	.60
New insertion of indwelling catheter CAT1	.02	.02	.00	.09	.71
Worsening bowel continence CNT2	.19	.09	.00	.41	.88
++Worsening bladder continence CNT3	.19	.09	.00	.49	.87
++Pain worsening PAN1	.10	.05	.00	.26	.73
++Worsening pressure ulcers PRU4	.07	.04	.00	.27	.74

Table 6.1

QI Rates and Weighted Kappas

Quality Indicator	QI Proportional Rate – The Average Across Facilities	Standard Deviation of the QI Rate	The Rate in the Facility with the Lowest Proportional Problem	The Rate in the Facility with the Highest Proportional Problem	Average Weighted Kappa for MDS Items Composing the QI [1]
Post-acute Prevalence					
++Failure to improve and manage delirium (pilot) DELX	**.03**	**.03**	**.00**	**.16**	**.65**
++Inadequate pain management (pilot) PAIX	**.27**	**.10**	**.02**	**.60**	**.72**
Post-acute Incidence					
Failure to improve during early post-acute period ADLX	.63	.19	.14	.99	.72
++Failure to improve bladder incontinence CNTX	.55	.09	.32	.79	.73
++Failure to prevent or improve pressure ulcers PRUX	.23	.09	.04	.50	.74
++Failure to improve or prevent respiratory problems RSPX	.92	.05	.77	.99	.53
++Improvement in Walking (pilot) WALX	**.28**	**.14**	**.03**	**.71**	**.77**

Table 6.1

QI Rates and Weighted Kappas

Quality Indicator	QI Proportional Rate – The Average Across Facilities	Standard Deviation of the QI Rate	The Rate in the Facility with the Lowest Proportional Problem	The Rate in the Facility with the Highest Proportional Problem	Average Weighted Kappa for MDS Items Composing the QI [1]

Notes:

2 Kappas below 0.4 reflect poor inter-rater reliability; a value between .40 and .60 is indicative of acceptable inter-assessor agreement; and a value of .75 or higher is indicative of superior inter-assessor reliability.

++ Quality indicator was risk-adjusted using facility admission profile.

* Validation analyses were not complete for these QIs.

BOLD items indicate measure was using in Nursing Home Quality Initiative Public Reporting Pilot

6.1.2 The Performance of all Prevalence-based Quality Indicators

Table 6.2 contains results for the average "percent agreement" between facility and research nurse assessors on 21 prevalence-based quality indicators. Overall level of agreement in the population of raters was high, with only the "Little or no activities" QI demonstrating lower than 70 percent agreement. Most QIs were near or above the 90 percent agreement mark. The "population average" Kappas are presented in Table 6.2 along with the "facility-specific average" Kappas for each QI. Again, only the "Little or no activities" QI performed poorly.

6.1.3 Other Analyses

We also examined the following issues in these analyses:

- The effect of elapsed time on reliability;
- Inter- and intra-state variation in reliability;
- A detailed case study of the reliability of the MDS items on pain (items J2a and J2b);
- An analysis of consistently poor-performing facilities (in terms of MDS reliability); and
- An assessment of the direction of measurement bias, using the Gamma statistic.

With the exception of the analysis of measurement bias, which may be found in Section 7.0, a discussion of the above-referenced analyses and findings may be found in Appendix H. In summary, the key findings regarding MDS and QI reliability are as follows:

- Overall level of agreement in the population of raters was high, with only the "Little or no activities" QI demonstrating lower than 70 percent agreement. Most QIs were near or above the 90 percent agreement mark;
- We found no significant differences between facility and research assessor level of agreement attributable to elapsed time between pairs of assessments;

- As seen in our previous research, we did find considerable variability of reliability statistics within and across states;
- Facilities tended to be either "good" raters or "poor" raters, as measured by Kappa scores with exceptional reliability (i.e. greater than .75) plotted against Kappa scores with poor reliability (i.e., Kappas below .40);
- Only a handful of facilities were found to be problematic based on the Kappa and Gamma analyses; and when validation models were tested with and without this small number of facilities there was little or no effect on the results. Thus, the findings that follow included all sampled facilities.

Table 6.2
Performance of 22 Quality Measures

Quality Measure	Percent Agreement	Population Average Kappa	Facility-Specific Average Kappa
Behavior high & low risk (chsra; beh01)	89.88	0.65	0.60
Behavior high risk (chsra; beh02)	86.89	0.65	0.63
Behavior low risk (chsra; beh03)	96.43	0.51	0.75
Little or no activities (chsra; soc02)	65.39	0.21	0.21
Catheter (chsra; cat02)	92.59	0.67	0.68
UTI (chsra; cnt04)	89.16	0.48	0.42
Incontinence hi & lo risk (chsra; cnt01)	91.47	0.83	0.79
Incontinence high risk (chsra; cnt05)	97.59	0.75	0.74
Incontinence low risk (chsra; cnt06)	90.76	0.8	0.76
Tube feeding (ramsey; nut01)	98.19	0.87	0.82
Low BMI (megaqi; bmi0x)	96.7	0.87	0.83
Weight loss (chsra; wgt1)	*	*	*
Infection flare-up (megaqi; inf0x)	79.67	0.45	0.37
Pain poorly managed (megaqi; pai0x)	86.57	0.57	0.50
PU high & low risk (chsra; pru01)	88.68	0.6	0.54
PU high risk (chsra; pru02)	85.33	0.61	0.58
PU low risk (chsra; pru03)	92.29	0.52	0.81
Burns abrasions bruises (megaqi; bur0x)	90.28	0.24	0.57
Restraints (chsra; res01)	91.39	0.53	0.51
Antipsych hi & low risk (chsra; drg01)	94.63	0.82	0.78
Antipsychotic high risk (chsra; drg02)	89.87	0.8	0.67
Antipsychotic low risk (chsra; drg03)	95.63	0.81	0.77

Note: *Analyses were not completed for this QI.

6.2 Primary Validation Findings

Appendix I & J display the results of the relationship between the a priori hypotheses and the quality indicators, and the findings are as anticipated. There are more positive, significant findings than one would have expected by chance alone. In fact the rate of observed findings of this type is over twice what one would have expected had the relationships been simply random. But, it is also true that for the typical QI, which had slightly less than 16 *a priori* hypotheses, only 2.4 of these hypotheses are found to be significant, and in the direction hypothesized (note, in the appendix tables, there are also a few shaded values, representing instances where the direction of the observed, significant relationship was counter to that which had been hypothesized). Nevertheless, the fact remains that these relationships do begin to lay the foundation for our validation rationale for a number of the quality indicators.

For the item-specific preventive and responsive analyses, many more positive findings are observed (see Appendix K, Tables Ka, Kb, Kc, and Kd). There is support for both the preventive and responsive hypotheses.

These individual findings are summarized in Table 6.3. The rows of the table reference the individual quality indicators, arranged as follows: chronic prevalence indicators, chronic incidence indicators, post-acute prevalence indicators, and post-acute incidence indicators. There are seven additional columns to the table. The first three present the count of significant, supportive validation elements for each quality indicator, with separate counts for the number that fall under the preventive and responsive domains, and a final count of the total number of supportive validation elements for the indicator. Columns 4 through 6 provide the Multiple R correlation estimate of the relationship between the pool of significant validation elements and the quality indicator. The last column on the table, labeled "Degree of Validity", provides the final assessment of the confidence one can have in the quality indicator at the end of this validation process. There are three possible classifications: Level I, Top validity, represents those quality indicators with the strongest support. Level II, Mid, represents the remainder of the validated indicators. Level III, Not Validated, represents indicators that failed to be supported in this analysis. In their current form, there is insufficient reason to believe that they provide a reasonable facility estimate for the quality problems they seek to address.

Let us walk through the first row of the table, for the Behavioral symptoms (high and low risk) prevalence indicator. In terms of the count of supportive elements, there are seven in total, three preventive and four responsive. The overall multiple R equals .43, and is .34 for the preventive elements and .31 for the responsive elements. Moving to the last column, the net result of this analysis supports the validity of this quality indicator. This is a Level II, Mid validation finding. As can be seen in the "notes" to Table 6.3, the criteria for a QI to be categorized as Level II requires a preventive Multiple R equal to or greater than .30 OR a total Multiple R equal to or greater than .40. This QI is found to be valid at the Level II category based upon both its total Multiple R and preventive Multiple R scores.

For all quality indicators the findings are as follows:

- Thirteen chronic quality indicators were at Level I, Top. Eight are prevalence indicators, while five are incidence indicators. There are eight clinical complexity indicators (Bladder/bowel incontinence prevalence in total and for the two risk subgroups, Infection prevalence, UTI prevalence, Inadequate pain management, Pressure ulcer prevalence, Worsening bladder continence); one service indicator (Catheter prevalence,); and four functional indicators (Late-loss ADL worsening, ADL worsening, locomotion worsening, improvement in walking).
- Sixteen chronic quality indicators are at Level II, Mid. Eight are prevalence indicators, while eight are incidence indicators.
- Seven chronic quality indicators are at Level III, NOT Validated. They include Behavior prevalence high and low risk (although the overall, or combined indicator was Level I), Weight loss, Antipsychotic use high and low (although the overall, or combined indicator was Level II), Worsening behavior, and Worsening pressure ulcers.
- Four post-acute care quality indicators were at Level I, Top; two were at Level II, Mid; and one (Failure to prevent or improve pressure ulcers) was at Level III, NOT Validated.

Table 6.3

Summary Measures of Quality Indicator Validity

Quality Indicator	Count of Significant Preventive Data Elements[1]	Count of Significant Responsive / Reactive Data Elements	Total Count of Significant Data Elements	Multiple R (Measure of Association) For Preventive Elements	Multiple R For Responsive Elements	Multiple R for All Elements	Degree of Validity [2] I TOP II MID III NOT Valid
Chronic Prevalence							
++Behavior symptoms (high&low risk) BEH1	3	4	7	.34	.31	.43	II
++Behavior symptoms (high risk) BEH2	1	3	4	.25	.30	.39	III
++Behavior symptoms (low risk) BEH3	0	0	0	--	--	--	III
Little or no activity SOC2	8	1	9	.39	.13	.44	II
Prevalence of indwelling catheter CAT2	5	6	11	.45	.71	.78	I
++Bladder/bowel incontinence (high&low risk) CNT1	7	3	10	.50	.45	.66	I
++Bladder/bowel incontinence (high risk) CNT5	8	2	10	.57	.35	.65	I
++Bladder/bowel incontinence (low risk) CNT6	5	3	8	.47	.31	.56	I
Urinary tract infections CNT4	7	8	15	.51	.41	.59	I
Falls FAL1	4	7	11	.27	.40	.50	II
++Infections (pilot) INFX	**6**	**9**	**15**	**.46**	**.36**	**.53**	**I**
++Feeding Tubes NUT1	7	8	15	.44	.40	.54	II
++Low Body Mass Index BMIX	6	1	7	.39	.20	.41	II
++Weight loss (pilot) WGT1	**3**	**0**	**3**	**.27**	**--**	**.27**	**III**
++Inadequate Pain Management (pilot) PAIX	**5**	**4**	**9**	**.32**	**.67**	**.74**	**I**

Table 6.3

Summary Measures of Quality Indicator Validity

Quality Indicator	Count of Significant Preventive Data Elements[1]	Count of Significant Responsive / Reactive Data Elements	Total Count of Significant Data Elements	Multiple R (Measure of Association) For Preventive Elements	Multiple R For Responsive Elements	Multiple R for All Elements	Degree of Validity [2] I TOP II MID III NOT Valid
++Pressure ulcers (high&low risk) (pilot) PRU1	**10**	**12**	**22**	**.48**	**.43**	**.59**	**I**
++Pressure ulcers (high risk) PRU2	*	*	*	*	*	*	*
++Pressure ulcers (low risk) PRU3	*	*	*	*	*	*	*
++Burns, skin tears or cuts BURX	4	7	11	.30	.34	.47	II
Restraints used daily (pilot) RES1	**3**	**7**	**10**	**.33**	**.48**	**.52**	**II**
++Antipsychotic use (high&low risk) (pilot) DRG1	**5**	**3**	**8**	**.32**	**.31**	**.47**	**II**
++Antipsychotic use (high risk) DRG2	0	1	1	--	.31	.31	III
++Antipsychotic use (low risk) DRG3	1	3	4	.15	.35	.38	III
Chronic Incidence							
Late-loss ADL worsening (pilot) ADL1	**13**	**1**	**14**	**.49**	**.26**	**.51**	**I**
ADL worsening ADL2	17	1	18	.57	.07	.57	I
ADL improvement ADL3	5	0	5	.39	--	.39	II
++Locomotion worsening MOB1	8	1	9	.62	.09	.62	I
++Improvement in walking WALX	9	0	9	.64	--	.64	I
++Cognition worsening COG1	12	8	20	.40	.34	.52	II
++Worsening communication COM1	3	5	8	.29	.31	.41	II
++Delirium DELX	10	0	10	.40	--	.40	II

Table 6.3

Summary Measures of Quality Indicator Validity

Quality Indicator	Count of Significant Preventive Data Elements[1]	Count of Significant Responsive / Reactive Data Elements	Total Count of Significant Data Elements	Multiple R (Measure of Association) For Preventive Elements	Multiple R For Responsive Elements	Multiple R for All Elements	Degree of Validity [2] I TOP II MID III NOT Valid
++Worsening behavior BEH4	1	1	2	.15	.17	.24	III
++Depressed anxious mood worsening MOD3	7	0	7	.31	--	.31	II
New insertion of indwelling catheter CAT1	8	6	14	.40	.24	.44	II
Worsening bowel continence CNT2	3	1	4	.25	.30	.45	II
++Worsening bladder continence CNT3	6	5	11	.39	.40	.63	I
++Pain worsening PAN1	10	5	15	.37	.40	.51	II
++Worsening pressure ulcers PRU4	3	2	5	.27	.23	.35	III
Post-acute Prevalence [3]							
++Failure to improve and manage delirium (pilot) DELX	**6**	**3**	**9**	**.58**	**.36**	**.62**	**I**
++Inadequate pain management (pilot) PAIX	**5**	**2**	**7**	**.52**	**.36**	**.64**	**I**
Post-acute Incidence							
Failure to improve during early post-acute period ADLX	9	0	9	.59	--	.59	I
++Failure to improve bladder incontinence CNTX	3	0	3	.37	--	.37	II
++Failure to prevent or improve pressure ulcers PRUX	1	0	1	.12	--	.12	III

Table 6.3

Summary Measures of Quality Indicator Validity

Quality Indicator	Count of Significant Preventive Data Elements[1]	Count of Significant Responsive / Reactive Data Elements	Total Count of Significant Data Elements	Multiple R (Measure of Association) For Preventive Elements	Multiple R For Responsive Elements	Multiple R for All Elements	Degree of Validity [2] I TOP II MID III NOT Valid
++Failure to improve or prevent respiratory problems RSPX	2	0	2	.42	--	.42	II
++Improvement in Walking (pilot) WALX	**4**	**0**	**4**	**.48**	**--**	**.48**	**I**

Notes:

[1] An alpha significance level for the correlation between the validation element and the quality indicator of .09 or lower.

[2] Level I -- Preventive Multiple R Equal to or Greater than .45 – OR -- Total Multiple R equal to or greater than .55

Level II -- Preventive Multiple R Equal to or Greater than .30 – OR -- Total Multiple R equal to or greater than .40

Level III -- Preventive Multiple R Less than .30 – OR -- Total Multiple R less than .40

[3] The sample utilized in evaluation of the post-acute care QIs includes hospital-based transitional care units (TCUs) only [maximum N = 52 facilities]. At the same time, we note that this was one of two analytic samples that could have been used to evaluate the post-acute indicators. Under a second sampling strategy, the TCU sample could be supplemented through the addition of 104 chronic nursing facilities. In each of these facilities there were sufficient numbers of Medicare residents on which to calculate the post-acute quality indicators. Had this second sample approach been the primary strategy to be followed, rather than the TCU approach on which this task rests, the Failure to Prevent or Improve Pressure Ulcer quality indicator would not have been rejected. In fact it would have been placed in Level I, the highest validation category. At the other extreme, had this alternative approach been used, the Improvement in Walking quality indicator would have been placed in Level III, Not Validated.

++ Quality indicator was risk-adjusted using facility admission profile.

* Validation analyses were not complete for these QIs.

--- Indicates that statistics could not be generated due to lack of significant data elements.

BOLD items indicate measure was using in Nursing Home Quality Initiative Public Reporting Pilot

7.0 Analysis of the Facility Admission Profile

7.1 Background

A main concern in the implementation of an indicator-based quality reporting system is that judgments based on those quality indicators (QIs) might be influenced by facility characteristics other than quality of care. In past work, this project team investigated the impact of casemix differences resulting from differential admission or discharge practices and of differential ascertainment as likely sources for such biased assessments. Our investigation, discussion previously in this report, confirms this concern. The specification of appropriate risk adjustment models is a key requirement for the validity of any QI. Prior analyses conducted revealed that, particularly in smaller facilities, rankings based on some QIs may vary substantially over time and, therefore, that statements about QI performance in smaller facilities cannot be made with much statistical confidence.

In attempts to capture these differential effects on quality indicator rankings, a series of analyses were conducted, resulting in the development of a new risk adjustment method that incorporates facility-admitting characteristics into the construction of QIs. We refer to this adjustment method as the "facility admission profile" (FAP). In prior work, the project team recommended the use of this facility-level adjuster on some but not all QIs. In general, use of the FAP was recommended for QIs where

1) the adjustment model performs well statistically,
2) the measurement of the quality dimension in question is subjective and more prone to the differential effects of assessment acumen or bias, and
3) the facility will encounter significant challenges to effecting change within a quality measure domain. For example, a FAP was suggested for the "Inadequate pain management" QI, where a facility's ability to impact on the condition is more challenging. On the other hand, facilities with a "restraint-free" philosophy, which would be in keeping with national trends, have the ability to limit, if not totally avoid, physical restraint use subsequent to resident admission. Thus, no FAP adjustment is recommended for the "prevalence of restraints" quality indicator.

7.1.1 Public Sentiment about the Facility Admission Profile (FAP)

There has been great debate about the issue of risk adjustment when making judgments about quality. Some stakeholders strongly advocate for risk adjustment, others argue strongly against. Arguments vary, but most advocates of risk adjustment believe that some adjustment of quality indicators is necessary to prevent biased rankings of facilities. Bias in quality ranking may be introduced when the quality measure does not sufficiently capture the variance in resident populations at given facilities. Opponents to risk adjustment of quality indicators argue that no adjustment is better than "over" adjustment, and that, in the absence of a perfect risk adjustment measure or method, indicators should remain unadjusted.

This project team believes strongly that risk should be taken into consideration in the measurement of quality, and we have been exploring risk adjustment techniques throughout this research process. As stated above, the FAP emerged as a response to this commitment. On a preliminary basis, we developed the FAP for use with the initially recommended set of existing and newly developed

quality indicators (pending results of this validation study). There has been great debate about this particular form of risk adjustment. Those concerned with this draft recommendation have argued that

- the FAP makes facilities that might be considered to look 'bad' or 'average' with an non FAP-adjusted QI, inappropriately look 'average' or 'good' with a FAP-adjusted QI;
- the FAP should not be applied universally to all quality indicators (a position that the research team is in agreement with);
- since the FAP (in the case of chronic care QIs) is based on an admission assessment that might not be conducted for up to 14 days after admission, the FAP can actually reflect the early effects of care (both positive and negative) provided by the facility. During this delay, many of the quality concerns measured by the QIs are likely to occur (that were not present upon arrival). Since these QIs will show up on the 'assessment', the FAP will adjust down the facility's QI rates (thereby making the facility look better then it should).
- there are also arguably incentives for facilities to report greater disability upon admission for both reimbursement and quality assurance purposes, which may skew the admission picture captured by the FAP, and
- the FAP is too complicated and therefore will not be credible to end users of the information (e.g., facility staff, consumers).

In response to these concerns, and in line with the long-standing plan for the national validation of the quality indicators, the project team undertook a series of analyses intended to evaluate the utility of the FAP as a risk adjustor for nursing facility quality indicators. Methods and findings regarding this work are reported here. We want to stress that these findings reflect preliminary work in a very complex area of inquiry. We will have greater confidence in our conclusions after further modeling, replicating at the national level some of the initial analyses that were conducted on a limited number of states.

7.2 Analyses Conducted to Assess Validity and Measurement Error

We report on two separate analyses. Each examined different aspects of the facility-level adjustment mechanism.

- First, we compared the validity of raw, or non FAP-adjusted, quality indicators to the validity of FAP-adjusted indicators.
- Second, we tested the impact of systematic measurement bias on quality indicators, as described below.

While each of these sets of analyses are still underway, the preliminary results obtained to date provide useful information about the performance of the FAP and should be considered as decisions are made regarding risk adjustment of publicly reported quality indicators.

7.2.1 Validation Findings for Non FAP-adjusted vs. FAP-adjusted Quality Indicators

We compared the results of the validation analyses with and without the FAP adjustment. Results of these comparisons are summarized in Table 7.1. Overall, there were relatively small differences in the quantitative results. That is, in most cases, the amount of variability accounted for by the

validation elements was of a comparable magnitude in the FAP and non FAP-adjusted forms of the QI. While in a number of instances the number of validation elements found to be related to the non-FAP form of the QIs was fewer than in the FAP-adjusted version, the statistical measures of model fit tended to be similar. Therefore, in most cases the level of validation for FAP and non FAP-adjusted forms of the QI was comparable, given our array of validation elements.

As can be seen in Table 7.1, FAP-adjusted QIs differed from non-FAP QIs in eight instances. In half of these instances, the FAP-adjusted QI indicated a higher level of validity than the non-FAP QI. In the other half, FAP adjustment implied a lower level of validity. However, it is worth mentioning that three of the four cases where the validation level was greater with FAP adjustment represent cases where a QI is classified as "Not Valid" without FAP, and "Top" or "Mid" level of validity *with* FAP adjustment. The four QIs where a higher level of validity was achieved with FAP adjustment included "Behavioral problems affecting others (high and low risk)", and three post-acute care quality indicators, "Improvement in walking", “Inadequate pain management” and "Failure to improve bladder incontinence." Only two of the four cases where a higher level of validity was observed without FAP adjustment represented a shift from a "Not Valid" level to a "Mid" or "Top" level ("Behavioral problems affecting others (high risk)” and “antipychotic use (low risk)”).

Overall, these comparative analyses demonstrate that, under the model used in our national validation of the quality indicators, there is little evidence that the FAP adjustment results in an array of QI scores that perform better than QIs without FAP adjustment in the statistical sense. The FAP models did not, as had been hoped, out perform the non FAP-adjusted models.

Table 7.1

Comparison of Validation Results with and without Adjustment for Facility Admission Profile (FAP)

Quality Indicators with FAP Adjustment	Count of Significant Preventive Elements [1]	Count of Significant Responsive Elements	Total Count of Significant Responsive and Preventive Elements	Multiple R (Measure of Association) for Preventive Elements	Multiple R (Measure of Association) for Responsive Elements	Multiple R for All Elements	Degree of Validity [2] I TOP II MID III NOT Valid
Chronic Prevalence							
Behavior high & low risk (chsra; beh01)							
FAP	3	4	7	0.34	0.31	0.43	II
without FAP	1	2	3	0.16	0.31	0.33	III
Behavior high risk (chsra; beh02)							
FAP	1	3	4	0.25	0.30	0.39	III
without FAP	0	2	2	---	0.38	0.40	II
Behavior low risk (chsra; beh03)							
FAP	0	0	0	---	---	---	III
without FAP	0	0	0	---	---	---	III
Incontinence hi & lo risk (chsra; cnt01)							
FAP	7	3	10	0.50	0.45	0.66	I
without FAP	6	2	8	0.52	0.59	0.76	I
Incontinence high risk (chsra; cnt05)							
FAP	8	2	10	0.57	0.35	0.65	I
without FAP	5	0	5	0.58	---	0.58	I
Incontinence low risk (chsra; cnt06)							
FAP	5	3	8	0.47	0.31	0.56	I
without FAP	4	2	6	0.50	0.44	0.65	I
Infection (megaqi; inf0x) (pilot)	**6**	**9**	**15**	**0.46**	**0.36**	**0.53**	**I**
FAP	**6**	**8**	**14**	**0.51**	**0.41**	**0.59**	**I**
without FAP							
Tube feeding (ramsey; nut01)							
FAP	7	8	15	0.44	0.40	0.54	II
without FAP	5	8	13	0.48	0.82	0.88	I
Low BMI (megaqi;bmi0x)							
FAP	6	1	7	0.39	0.20	0.41	II
without FAP	5	1	6	0.37	0.19	0.39	II

Table 7.1

Comparison of Validation Results with and without Adjustment for Facility Admission Profile (FAP)

Quality Indicators with FAP Adjustment	Count of Significant Preventive Elements [1]	Count of Significant Responsive Elements	Total Count of Significant Responsive and Preventive Elements	Multiple R (Measure of Association) for Preventive Elements	Multiple R (Measure of Association) for Responsive Elements	Multiple R for All Elements	Degree of Validity [2] I TOP II MID III NOT Valid
Weight loss (ltcq; wgt01) (pilot)							
FAP	**3**	**0**	**3**	**0.27**	**---**	**0.27**	**III**
without FAP	**3**	**0**	**3**	**0.26**	**---**	**0.26**	**III**
Pain poorly managed (megaqi; pai0x) (pilot)							
FAP	**5**	**4**	**9**	**0.32**	**0.67**	**0.74**	**I**
without FAP	**2**	**2**	**4**	**0.26**	**0.78**	**0.82**	**I**
PU high & low risk (chsra; pru01) (pilot)	**10**	**12**	**22**	**0.48**	**0.43**	**0.59**	**I**
FAP	**7**	**12**	**19**	**0.47**	**0.43**	**0.58**	**I**
without FAP							
PU low risk (chsra; pru02)							
FAP	*	*	*	*	*	*	*
without FAP	*	*	*	*	*	*	*
PU high risk (chsra; pru03) [2]							
FAP	*	*	*	*	*	*	*
without FAP	*	*	*	*	*	*	*
Burns abrasions bruises (megaqi; bur0x)							
FAP	4	7	11	0.30	0.34	0.47	II
without FAP	3	7	10	0.32	0.38	0.52	II
Antipsychotic high & low risk (chsra; drg01) (pilot)							
FAP	**5**	**3**	**8**	**0.32**	**0.31**	**0.47**	**II**
without FAP	**4**	**3**	**7**	**0.29**	**0.52**	**0.62**	**I**
Antipsychotic high risk (chsra; drg02)							
FAP	0	1	1	---	0.31	0.31	III
without FAP	0	0	0	---	---	---	III

Table 7.1

Comparison of Validation Results with and without Adjustment for Facility Admission Profile (FAP)

Quality Indicators with FAP Adjustment	Count of Significant Preventive Elements [1]	Count of Significant Responsive Elements	Total Count of Significant Responsive and Preventive Elements	Multiple R (Measure of Association) for Preventive Elements	Multiple R (Measure of Association) for Responsive Elements	Multiple R for All Elements	Degree of Validity [2] I TOP II MID III NOT Valid
Antipsychotic low risk (chsra; drg03)							
FAP	1	3	4	0.15	0.35	0.38	III
without FAP	0	3	3	---	0.51	0.51	II
Chronic Incidence							
Mobility decline (ltcq; mob01)							
FAP	8	1	9	0.62	0.09	0.62	I
without FAP	7	0	7	0.67	---	0.67	I
Walking performance (megaqi; wal0x)							
FAP	9	0	9	0.64	---	0.64	I
without FAP	8	0	8	0.67	---	0.67	I
Cognition worsening (ltcq; cog01)	12	8	20	0.40	0.34	0.52	II
FAP	12	8	20	0.39	0.34	0.52	II
without FAP							
Communication worsening (ltcq; com01)							
FAP	3	5	8	0.29	0.31	0.41	II
without FAP	3	5	8	0.28	0.32	0.42	II
Delirium not remitting (megaqi; del0x)							
FAP	10	0	10	0.40	---	0.40	II
without FAP	9	0	9	0.39	---	0.39	II
Behavior worsening (ltcq; beh04)							
FAP	1	1	2	0.15	0.17	0.24	III
without FAP	1	1	2	0.13	0.21	0.26	III
Depression new or worse (ltcq; mod03)							
FAP	7	0	7	0.31	---	0.31	II
without FAP	5	0	5	0.31	---	0.31	II

Table 7.1

Comparison of Validation Results with and without Adjustment for Facility Admission Profile (FAP)

Quality Indicators with FAP Adjustment	Count of Significant Preventive Elements [1]	Count of Significant Responsive Elements	Total Count of Significant Responsive and Preventive Elements	Multiple R (Measure of Association) for Preventive Elements	Multiple R (Measure of Association) for Responsive Elements	Multiple R for All Elements	Degree of Validity [2] I TOP II MID III NOT Valid
Bladder incont decline (ltcq; cnt03)							
FAP	6	5	11	0.39	0.40	0.63	I
without FAP	5	3	8	0.39	0.42	0.66	I
Pain worsening (ltcq; pan01)							
FAP	10	5	15	0.37	0.40	0.51	II
without FAP	8	5	13	0.39	0.37	0.49	II
PU onset or worsening (ltcq; pru04)							
FAP	3	2	5	0.27	0.23	0.35	III
without FAP	2	2	4	0.24	0.23	0.33	III
Post-acute Prevalence [3]							
Failure to improve or manage delirium (megaqi; del0x) (pilot)							
FAP	**6**	**3**	**9**	**0.58**	**0.36**	**0.62**	**I**
without FAP	**5**	**3**	**8**	**0.53**	**0.38**	**0.59**	**I**
Inadequate pain management (megaqi; pai0x) (pilot)							
FAP	**5**	**2**	**7**	**0.52**	**0.36**	**0.64**	**I**
without FAP	**0**	**0**	**0**	**---**	**---**	**---**	**III**
Post-acute Incidence [3]							
Failure to improve bladder incontinence (megaqi; cnt0x)							
FAP	3	0	3	0.37	---	0.37	II
without FAP	2	0	2	0.29	---	0.29	III

Table 7.1

Comparison of Validation Results with and without Adjustment for Facility Admission Profile (FAP)

Quality Indicators with FAP Adjustment	Count of Significant Preventive Elements [1]	Count of Significant Responsive Elements	Total Count of Significant Responsive and Preventive Elements	Multiple R (Measure of Association) for Preventive Elements	Multiple R (Measure of Association) for Responsive Elements	Multiple R for All Elements	Degree of Validity [2] I TOP II MID III NOT Valid
Failure to prevent/ improve PU (megaqi; pru0x)							
FAP	1	0	1	0.12	---	0.12	III
without FAP	1	0	1	0.24	---	0.24	III
Failure to improve/ prevent respiratory problems (megaqi; rsp0x)							
FAP	2	0	2	0.42	---	0.42	II
without FAP	1	0	1	0.38	---	0.38	II
Improvement in walking (megaqi; wal0x) (pilot)							
FAP	**4**	**0**	**4**	**0.48**	**---**	**0.48**	**I**
without FAP	**0**	**0**	**0**	**---**	**---**	**---**	**III**

Notes:

[1] An alpha significance level for the correlation between the validation element and the quality indicator of .09 or lower.

[2] Level I -- Preventive Multiple R Equal to or Greater than .45 – OR -- Total Multiple R equal to or greater than .55

Level II -- Preventive Multiple R Equal to or Greater than .30 – OR -- Total Multiple R equal to or greater than .40

Level III -- Preventive Multiple R Less than .30 – OR -- Total Multiple R less than .40

[3] The sample utilized in evaluation of the post-acute care QIs includes hospital-based transitional care units (TCUs) only [maximum N = 52 facilities]. At the same time, we note that this was one of two analytic samples that could have been used to evaluate the post-acute indicators. Under a second sampling strategy, the TCU sample could be supplemented through the addition of 104 chronic nursing facilities. In each of these facilities there were sufficient numbers of Medicare residents on which to calculate the post-acute quality indicators. Had this second sample approach been the primary strategy to be followed, rather than the TCU approach on which this task rests, the Failure to Prevent or Improve Pressure Ulcer quality indicator would not have been rejected. In fact it would have been placed in Level I, the highest validation category. At the other extreme, had this alternative approach been used, the Improvement in Walking quality indicator would have been placed in Level III, Not Validated.

* Validation analyses were not complete for these QIs.

--- Indicates that statistics could not be generated due to lack of significant data elements.

BOLD items indicate measure was using in Nursing Home Quality Initiative Public Reporting Pilot

7.2.2 Analysis of the Effect of Systematic Measurement Bias on the QI

One reason for examining the prevalence of systematic measurement bias in the QIs was because of concerns regarding inter-facility variation in the comprehensiveness of assessments. Cursory assessments might yield lower rates of clinical problems. Indeed, this concern was one of the principle motivations for the creation of the FAP that characterizes all residents admitted to a facility over the year prior to the measurement of the QI. Conceptually, the FAP has the potential of capturing the propensity of facility assessors to detect clinical problems that they inherited from the admitting location.

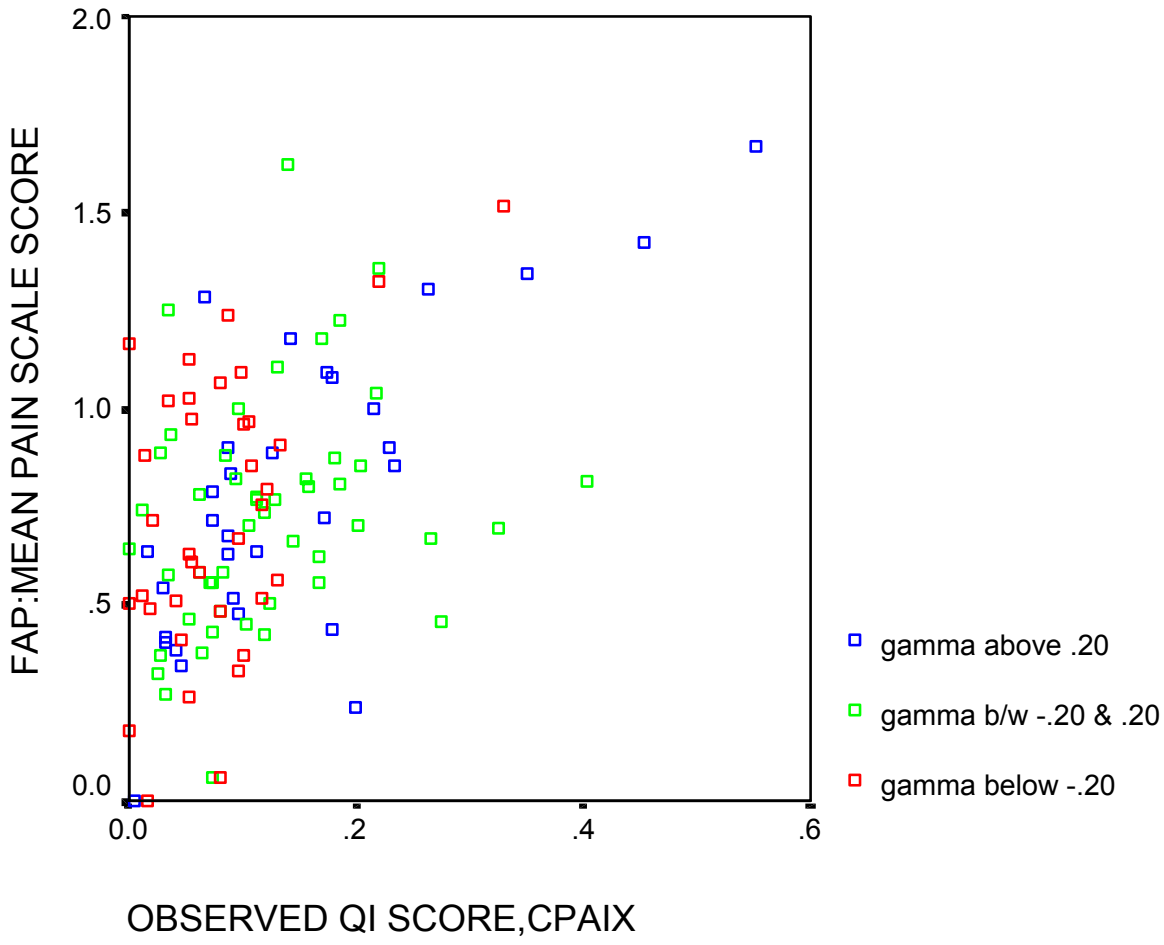
The Gamma statistic, which has been described in Section 5.3, provides a means for assessing the extent to which there is a directional bias in the disagreements between the research nurse assessors and the facility assessors. As noted earlier, facilities with high positive Gamma statistics were those that observed more clinical problems in a given domain than did our research nurses. Facilities with high negative Gamma statistics were less likely to observe clinical problems among residents than were the research nurses. Nonetheless, by and large there were relatively few facilities that consistently manifested high or low Gamma statistics on numerous measures, and the modal facility had Gamma statistics within +.2 and -.2 of the unbiased zero (0). Consequently, in conducting analyses of the effect of the Gamma on the QI measurement with and without the FAP, we used this cut-off to classify facilities as over or under reporting (or assessing) the clinical condition in question in the QI.

We anticipated that facilities with high negative Gamma statistics (facilities less likely to detect a clinical problem than were the research nurse assessors) would also have a FAP that would be correlated to the related QI. To that end, we examined whether the presence of high positive or high negative Gamma statistics attenuated the observed relationship between the QI and its associated FAP. This was done using both graphical means and multiple linear regression analysis. A complete description of the analyses applied to the "Inadequate pain management" QI is presented below followed by parallel analyses performed for the "Infection" QI.

Each nursing home's facility admission profile for pain and its observed QI score for pain are portrayed in the scatter plot below (Figure 1). The plot is further identified using Gamma values classified into 3 groups: Gamma above .20, Gamma between -.20 and .20, and Gamma below -.20.

As expected, facilities with the highest observed pain QI scores tended to have the highest Gamma scores. There are only a few "below -.20" facilities with observed pain QI scores above 20 percent. Conversely, many of the facilities with a Gamma score above .2 had the highest pain QI measures. On the other hand, there is little apparent pattern to the relationship between the three Gamma classes and the FAP score. That is, facilities with Gamma statistics in excess of .2 or less than -.2 were equally likely to have a FAP scale score under .5 and over 1.0. This suggests that the direction of bias in the measurement of the MDS items that make up the pain QI is not particularly related to the prevalence of pain among residents assessed at the time of their admission.

Figure 1. Scatter plot of Pain QI and Pain FAP, by Gamma Level Classification



To more formally test the effect of this Gamma class construct on the statistical relationship between the QI and the FAP, the observed pain QI score was regressed on the pain FAP to identify the association between the two items. Table 7.2a below shows the strong relationship between admission prevalence and the observed QI score in this sample of facilities participating in the validation study and serving chronic patients. The correlation between the two variables is about .5**.** Table 7.2b contains the results of the model after introducing two dummy variables to reflect the relative position of each facility's Gamma value. "Above .20" and "Below -.20" can be interpreted in reference to "between –.20 and .20" (omitted). The results reveal that introducing the two "dummy" Gamma values only modestly attenuates the relationship between the FAP and the observed QI. While, as hypothesized, a negative Gamma is significantly related to a facility's QI (t= -2.105; p=.038), the residual relationship between the QI and the FAP is not terribly different. Without the Gamma indicators in the model, the correlation between the FAP and QI is .5; with them included it drops to .45.

Table 7.2a
Observed Pain QI Regressed on Pain FAP

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.502[a]	.252	.246	.10865

a. Predictors: (Constant), FAP:MEAN PAIN SCALE SCORE

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-1.364E-02	.024		-.567	.572
	FAP:MEAN PAIN SCALE SCORE	.182	.029	.502	6.230	.000

a. Dependent Variable: OBSERVED QI SCORE,CPAIX

Table 7.2b
Observed Pain QI Regressed on FAP with Gamma Dummies

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.553[a]	.305	.287	.10484

a. Predictors: (Constant), FAP:MEAN PAIN SCALE SCORE, Above .20, Below -.20

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1.487E-03	.026		.057	.954
	FAP:MEAN PAIN SCALE SCORE	.167	.029	.457	5.755	.000
	Gamma above .20	3.637E-02	.024	.132	1.519	.132
	Gamma below -.20	-4.886E-02	.023	-.181	-2.105	.038

a. Dependent Variable: OBSERVED QI SCORE,CPAIX

A similar analysis was performed for the "Infection" QI. The results are presented in Figure 2. Relatively few facilities have Gamma values exceeding .20 for the infection QI and those facilities with high Gamma values appear clustered along the diagonal of the relationship between the observed QI and the FAP. The infection FAP is clearly associated with the observed Infection QI (Table 7.3a), but analysis failed to detect any attenuation of the relationship after introducing the Gamma-dummied values (Table 7.3b).

Figure 2
Scatter plot of Infection QI and Infection FAP, by Gamma Classification

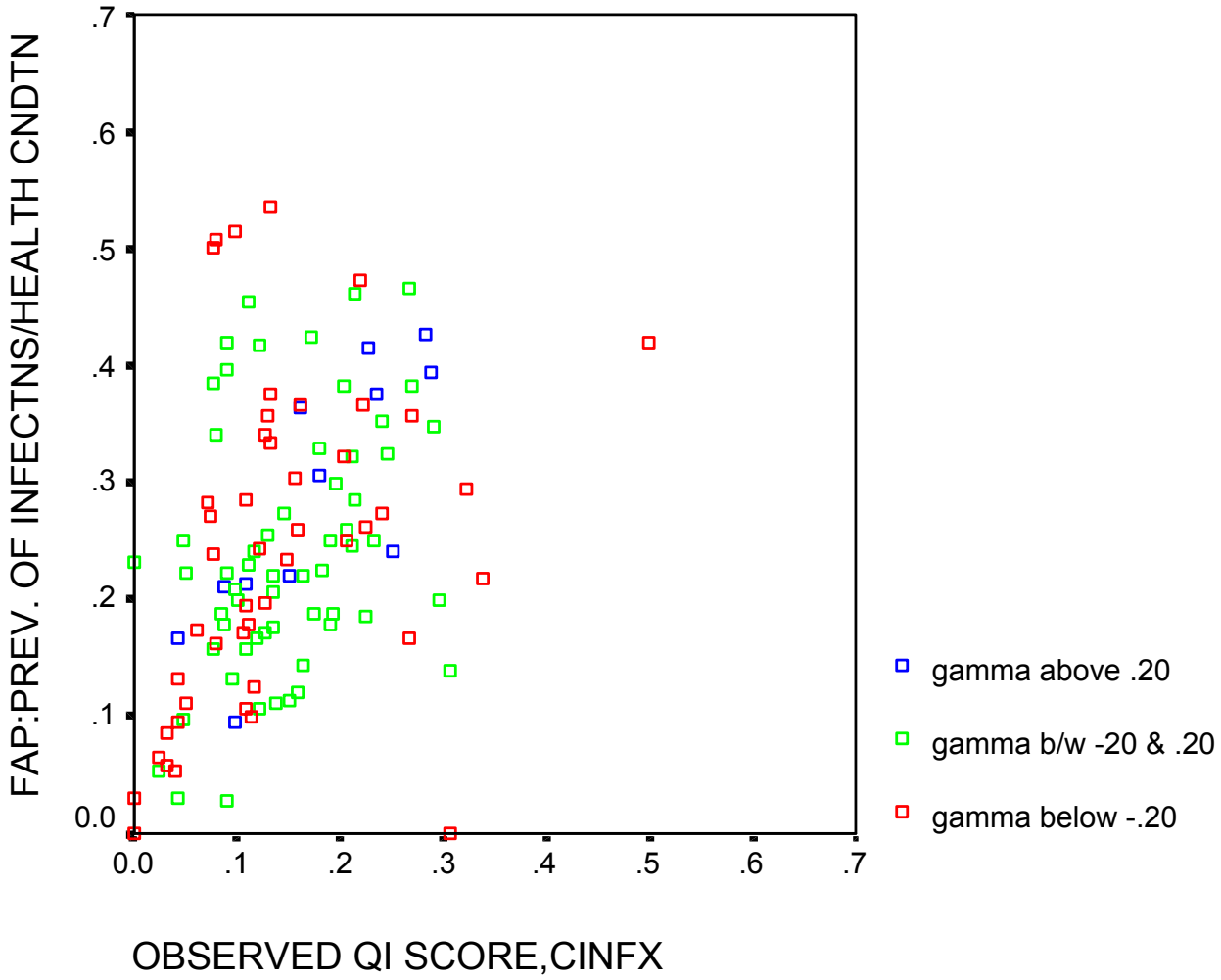


Table 7.3a
Observed Infection QI Regressed on Infection FAP

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.453[a]	.206	.199	.14063

a. Predictors: (Constant), FAP:PREVALENCE OF INFECTNS/HEALTH CNDTN

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.889E-02	.028		1.025	.307
	FAP:PREVALENCE OF INFECTNS/HEALTH CNDTN	.552	.100	.453	5.549	.000

a. Dependent Variable: OBSERVED QI SCORE,CINFX

Table 7.3b
Observed Infection QI Regressed on FAP with Gamma Dummies

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.489[a]	.239	.219	.13954

a. Predictors: (Constant) FAP:PREVALENCE OF INFECTNS/HEALTH CNDTN, Above .20, Below -.20

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.215E-02	.031		.722	.471
	FAP:PREVALENCE OF INFECTNS/HEALTH CNDTN	.579	.102	.468	5.665	.000
	Gamma above .20	3.448E-02	.043	.068	.796	.428
	Gamma below -20	-1.374E-02	.027	-.043	-.502	.616

a. Dependent Variable: OBSERVED QI SCORE,CINFX

Similar results were obtained for pressure ulcers, behavior problems and bladder incontinence (additional tables can be made available upon request). In each case, the introduction of the dummy variables for the Gamma value exceeding +/- .20 failed to significantly attenuate the observed relationship between the FAP and the QI. These results suggest that the FAP cannot be considered an adequate and robust measure of ascertainment bias. Rather, to the extent that our Gamma statistic measures the presence of directional measurement bias, it appears to be weakly, but independently (if at all) associated with the QI in a way that does not meaningfully affect the relationship between the FAP and the QI.

Future analyses could be designed to re-examine these data to determine whether the relative ranking of the facilities on the various QIs is altered when FAP-adjusted and non FAP-adjusted versions are used and when adjustment is based solely on the Gamma statistic that is available for each of the facilities participating in the validation study. We anticipate that, as we have seen in comparing FAP-adjusted and non FAP-adjusted data from all US nursing facilities, facilities' rankings may change, with more facilities scored near the median of the QI distribution. In light of the statistical and graphical analyses presented above, we anticipate that adjusting the facility QI distribution only with the Gamma statistic measured for each facility will not materially affect the distribution. Nonetheless, it is the next logical step in the analysis.

8.0 Conclusions, Recommendations and Next Steps

8.1 Conclusions and Recommendations Regarding the Validity of These Quality Indicators

In this national validation study, there is strong evidence that many of the set of 45 reviewed quality indicators capture meaningful aspects of nursing facility performance, and are reliably measured. We highly recommend for use by CMS and nursing facilities any of the QIs that fall into the Level I validation category, as these QIs have the strongest degree of evidence that they represent real care processes in nursing facilities. The chronic care quality indicators with the highest level of validity include:

- Prevalence of indwelling catheter;
- Bladder/bowel incontinence (high and low risk, high risk, low risk);
- Urinary tract infections;
- Infections;
- Inadequate pain management;
- Pressure ulcers (high and low risk);
- Late-loss ADL worsening;
- ADL worsening;
- Locomotion worsening;
- Improvement in walking; and
- Worsening bladder continence.

Four post-acute care quality indicators are highly valid, including:

- Failure to improve and manage delirium[10];
- Inadequate pain management;
- Failure to improve during early post-acute period; and
- Improvement in walking.

The chronic quality indicators that we recommend rejecting for further use at this time are:

- Behavior symptoms (high risk and low risk);
- Weight loss;
- Antipsychotic use (high risk and low risk);
- Worsening behavior; and
- Worsening pressure ulcers.

The post-acute care indicator that proved not to be valid is "Failure to Prevent or Improve Pressure Ulcers" and therefore should be rejected for use by CMS.

[10] Again, this QI has a very low rate of occurrence (three percent) in our study sample. The national distribution of this indicator should be examined as CMS makes a final determination as to this QI's overall utility.

Those QIs that fall into the Level II – Mid Valid category are deemed appropriate for use in measuring nursing facility quality, as they do offer evidence of validity; they are simply not as highly recommended to CMS as those QIs falling into the "Top" (Level I) validation category. In making final determinations about the utility of these QIs for performance improvement, public reporting or other purposes, CMS may want to review both the prevalence and the reliability of these indicators.

A special note is warranted on the "Little or No Activity" quality indicator. While based on the validation effort it was judged to fall into the Mid-Valid (Level II) category, the MDS item on which the indicator is based was found to have poor reliability. Should CMS choose to utilize this indicator for public reporting, facilities will need instruction on proper coding of this assessment item.

In addition to determining which of these sets of nursing facility quality indicators are "valid", or reflecting the care outcomes and issues they are purported to reflect, these results provide evidence that quality indicators measure aspects of care quality that may be amenable to modification through facility practice. For example, facility staffing and policies, practices or procedures are found to be related to resident quality outcomes and therefore may be modified by facilities to enhance quality of care delivery.

8.1.1 Conclusions Regarding the Validity and Utility of the Facility Admission Profile Method of Risk Adjustment

At this time, the Project Team does not recommend the FAP for broad scale application as currently operationalized. From the series of three analyses described in this chapter, we find that

- Non FAP-adjusted and FAP-adjusted quality indicators were equally valid in all but eight instances. In four, two of which ("Improvement in walking – PAC" and "Inadequate pain management – PAC") are currently in the CMS Nursing Home Quality Initiative pilot project, validity was higher for the FAP-adjusted measures. For the other four, validity for the FAP-adjusted measures was lower. The FAP models did not out-perform the non-FAP models; they did not provide scores that were systematically superior.

- There is no evidence of systematic bias in facility reporting of the set of prevalence-based QIs evaluated here: the FAP therefore cannot be considered an adequate and robust measure of ascertainment bias.

In light of these findings, each of which might require additional work, we find no reason to continue to support the universal application of the FAP as currently operationalized. Nonetheless, our analyses also suggest that there are very real inter-facility differences in the mix of residents admitted and who remain to be served by the facility and that these differences are related to the distribution of facilities as measured by the non FAP-adjusted QIs as well as those relying only upon resident-level adjustment. Thus, we feel that additional research focusing on the testing of alternate resident- and facility-level adjustment variables is needed.

8.2 Next Steps

Much additional research undertaken by this project team is not presented here, as some remains preliminary and some issues are still under evaluation. One necessary next step in this process of making final recommendations to CMS about the utility of this set of quality indicators is to continue

work on exploring alternatives to the facility admission profile. Composite measures, such as a "proximity to death" index or a casemix index score, appear promising as alternative risk adjustors. These measures should be further conceptualized, and then modeled against the national MDS dataset to determine their performance and potential utility.

References

Abt Associates Inc. Identification and Evaluation of Existing Long-term and Post-acute Quality Indicators. October 2001.

Abt Associates Inc. Preliminary Report: Pilot Field Data Collection Efforts to Validate Nursing Home Quality Indicators. September 2001.

Banerjee, M., Capozzoli, L., McSweeney, L. and Sinha, D. (1999). Beyond Kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics*, 27: 3-23.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurements*. 20: 37-46.

Donabedian, A. *Explorations in Quality Assessment and Monitoring: The Definitions of Quality and Approaches to Its Assessment*, vol. 1. Ann Arbor, MI: Health Administration Press, 1980.

Flacker, J.M. and Kiely, D.K. A practical approach to identifying mortality-related factors in established long-term care residents. *Journal of the American Geriatrics Society*: 46:1012-5. 1998

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33: 159-174.

Morris, J.N., Murphy, K. and Nonemaker, S. Long Term Care Facility Resident Assessment Instrument User's Manual. October 1995.

Mukamel, D.B. and Brower, C.A. "The influence of risk adjustment methods on conclusions about quality of care in nursing homes based on outcome measures." *Gerontologist*. 38(6):695-703. December 1998.

Porell, F. and Caro, F.G. "Facility-level outcome performance measures for nursing homes." *Gerontologist.* 38(6)-665-83; December 1998.

Ramsay, J., Sainfort, F. and Zimmerman, D. (1995) "An Empirical Test of the Structure, Process, and Outcome Quality Paradigm Using Resident-Based, Nursing Facility Assessment Data." *American Journal of Medical Quality*. (10)2:63-75.

Sainfort, F., Ramsay, J. and Monato, H. (1995) "Conceptual and Methodological Sources of Variation in the Measurement of Nursing Facility Quality: An Evaluation of 24 Models and an Empirical Study." *Medical Care Research and Review.* (52)1:60-87.

Zimmerman, D. and Karon, S. "Measuring the Quality of Nursing Home Care: Risk Adjustment, Validation, and other 'Trivial' Issues." Presentation to the National Case Mix Reimbursement and Quality Assurance Conference. 22 September 1997.

Zimmerman, D. Karon, S., Arling, G. et al. (1995). "Development and Testing of Nursing Home QIs." *Health Care Financing Review*. 16(4):107-127.

Zimmerman, D., Karon, S., Swearengen, J. (1999). "The Quality Component of the Demonstration," in *DRAFT: CMS NHCMQ Demonstration Final Report*, March.